Moderators of the Internal Consistency of Error-Related Negativity Scores:

A Meta-Analysis of Internal Consistency Estimates

Peter E. Clayson*[1]

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

**Corresponding author at*: Peter Clayson, Department of Psychology, University of South Florida, 4202 E. Fowler Ave., PCD 4110, Tampa, FL, 33620-7200; *email*: clayson@usf.edu

Abstract

To ensure adequate reliability (i.e., internal consistency), it is common in studies using event-related brain potentials (ERPs) to exclude participants for having too few trials. This practice is particularly relevant for error-related ERPs, such as error-related negativity (ERN), where the number of recorded ERN trials is not entirely under the researcher's control. Furthermore, there is a widespread practice of inferring reliability based on published psychometric research, which assumes that internal consistency is a universal property of ERN. The present, preregistered reliability generalization study examined whether there is heterogeneity in internal consistency estimates of ERN scores and whether contextual factors moderate reliability. A total of 189 internal consistency estimates from 68 samples nested within 43 studies (*n* = 4,499 total participants) were analyzed. There was substantial heterogeneity in ERN score internal consistency, which was partially moderated by the type of paradigm (e.g., Stroop, flanker), the clinical status of the sample, the ocular artifact correction procedure, measurement sensors (single vs. cluster), and the approach to scoring and estimating reliability, suggesting that contextual factors impact internal consistency at the individual study level. Age, sex, year of publication, artifact rejection procedure, acquisition system, sample type (undergraduate vs. community), and length of mean amplitude window did not significantly moderate reliability. Notably, the overall estimated reliability of ERN scores was below established standards. Recommendations for improving ERN score reliability are provided, but the routine failure of most ERN studies to report internal consistency represents a substantial barrier to understanding the factors that impact reliability.

Keywords: error-related negativity (ERN), meta-analysis, reliability generalization, event-related potentials (ERPs), internal consistency, psychometrics

1. Introduction

The National Institute of Mental Health's Research Domain Criteria (RDoC) initiative emphasizes the study of dimensional conceptions of hypothetical constructs (Cuthbert & Insel, 2013; Cuthbert & Kozak, 2013; Kozak & Cuthbert, 2016). This initiative has focused attention on whether measurements of dimensional constructs show adequate psychometric properties for such endeavors. Recent findings indicate that some common, robust psychophysiological measurements of group or condition differences actually show poor internal consistency reliability (e.g., Fröhner, Teckentrup, Smolka, & Kroemer, 2019; Infantolino, Luking, Sauder, Curtin, & Hajcak, 2018). This poor internal consistency substantially limits their utility as dimensional measures for RDoC-inspired research, because the internal consistency of measurements is closely related to how well measurements can differentiate among participants. Hence, measures with poor internal consistency are poorly suited for studying individual differences. To ensure that psychophysiological measurements demonstrate adequate internal consistency, internal consistency should be routinely reported (Clayson & Miller, 2017b; Hajcak, Meyer, & Kotov, 2017; Infantolino et al., 2018; Thigpen, Kappenman, & Keil, 2017). In fact, *Psychophysiology* and *International Journal of Psychophysiology* recently adopted guidelines for reporting the internal consistency of measurements when examining individual differences (e.g., dimensional constructs).

An important reason that internal consistency needs to be routinely reported is that reliability is a property of scores in a given context, not a property of a measure (Thompson, 2003; Vacha-Haase, 1998). In studies of event-related brain potentials (ERPs), there is a widespread practice of inferring reliability based on published psychometric information, but this is based on the incorrect assumption that reliability is a stable property of an ERP. This practice is common in ERP studies, such as when adequate internal consistency is assumed when the number of trials retained for averaging satisfies a trial threshold for data inclusion from a previous psychometric analysis. Internal consistency estimates of an ERP reflect the stability of single-trial measurements within an individual (i.e., within-person variability) and the capability of measurements to distinguish between individual-participant measurements (i.e., between-person variability). For example, studies of the error-related negativity (ERN) often exclude participants with fewer than six to eight trials (Olvet & Hajcak, 2009) or fourteen trials (Larson, Baldwin, Good, & Fair, 2010) based on these studies of ERN score internal consistency. This practice represents a failure to appreciate the many contextual factors, such as sample characteristics and EEG data reduction parameters, that can influence ERP score reliability (Clayson & Miller, 2017b). The purpose of the present, preregistered study was to assess the utility of ERN as an individual-difference measure in healthy and clinical populations and to identify the relevant characteristics that influence internal consistency estimates.

Although excluding participants to achieve adequate internal consistency is common in ERP studies, this practice is of particular relevance for error-related ERPs, such as ERN. ERN is a negative deflection in the scalp-recorded ERP that occurs within 100 ms of error commission and indexes early error detection (Falkenstein, Hohnsbein, Hoormann, & Banke, 1991; Larson, Clayson, & Clawson, 2014b; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001). Although ERN follows error commission, the commission of errors is not entirely determined by the design of the experimental paradigm. For example, some high-performing participants commit relatively few errors, and these participants are often excluded from analyses for having too few error trials to achieve adequate internal consistency. Other participants with noisier data might have many trials excluded during artifact rejection that prevents all of the participants’ data from being included in participant averages, and these participants are also often excluded from analysis. The trial cutoffs for data exclusion are typically based on previous psychometric studies, and the number of trials retained for averaging is commonly used as a proxy for justifying adequate internal consistency of the recorded data.

There are many psychometric studies that examine trial cutoffs for obtaining adequate ERN internal consistency. However, these trial cutoffs vary considerably, and there is no universally applicable cutoff for obtaining adequate ERN score internal consistency. Recommendations range from 2 (Steele et al., 2016) to 15 trials (Fischer, Klein, & Ullsperger, 2017) in studies of healthy undergraduates. In clinical samples, recommendations range from 14 trials for psychotic disorders (Foti, Kotov, & Hajcak, 2013) to 23 trials for major depressive disorder (Baldwin, Larson, & Clayson, 2015) and up to 41 trials for anxiety disorders (Baldwin et al., 2015). In a study that compared ERN internal consistency across paradigms, trial-cutoff recommendations were 8 trials for the flanker task, 12 trials for the Go/NoGo task, and 18 trials for a Stroop task (Meyer, Riesel, & Hajcak, 2013). These studies highlight some of the contextual factors (e.g., sample and task) that can influence ERN score reliability by showing the number of trials needed to obtain a minimum internal consistency threshold, and the observation of different estimates underscores the importance of evaluating internal consistency on a study-by-study basis. Taken together, adequate internal consistency cannot be assumed based on the implementation of a trial cutoff from a previous psychometric analysis (Clayson & Miller, 2017b; Hajcak et al., 2017; Infantolino et al., 2018; Thigpen et al., 2017). Furthermore, there is a need for a synthesis of these disparate internal consistency estimates across different populations and paradigms to aid researchers during the planning stages of ERN studies.

Evaluating the score reliability of measurements, such as ERN, should be the first step of any experiment, because unreliable data can dramatically impact the results of a study. For example, unreliable scores can lead to magnitude or sign errors in between-group relationships (Flegal, Kit, & Graubard, 2017; Gelman & Carlin, 2014), reduced statistical power (Boudewyn, Luck, Farrens, & Kappenman, 2017; Clayson & Miller, 2017b; Fischer et al., 2017; Kolossa & Kopp, 2018; Luck & Gaspelin, 2017), and failures to find replicable effects (Cooper, Gonthier, Barch, & Braver, 2017; Loken & Gelman, 2017; Thigpen et al., 2017). In RDoC-inspired studies of individual differences or dimensional constructs, these problems with unreliable data are exacerbated because of larger within-person variability than between-person/within-group variability of measurements (Cooper et al., 2017; Fisher, Medaglia, & Jeronimus, 2018; Hedge, Powell, & Sumner, 2017; Loken & Gelman, 2017; Rouder & Haaf, 2018; Seghier & Price, 2018). All of these studies emphasize that the use of unreliable scores calls into question the statistical conclusions of a study.

Given the importance of score reliability, understanding the contextual factors that improve or weaken reliability is critical for optimizing ERP paradigms for basic and applied research. To identify such contextual factors, the present study used a reliability generalization analysis. A reliability generalization analysis synthesizes the reliability of scores across different applications of a measure (Botella, Suero, & Gambara, 2010; Thompson, 2003; Vacha-Haase, 1998). This meta-analytic technique assesses the heterogeneity of score reliability and identifies potential sources of this variance across samples. A synthesis of the ERN score reliability literature can inform future studies of potential sources of measurement error and provide guidance for optimizing measurement approaches for a particular application. An advantage of meta-analytic approaches is the capability to pool information across many studies to identify patterns of effects, and this is particularly advantageous for ERP studies which generally include few participants (see Clayson, Carbine, Baldwin, & Larson, 2019). In short, a reliability generalization analysis is well suited for determining the generalizability of ERN score internal consistency and the contextual factors that impact it.

The present, preregistered reliability generalization study had three aims. The first aim was to determine whether there is heterogeneity in ERN score internal consistency across samples and studies, and it was predicted that there would be significant heterogeneity. The second aim was to determine the influence of three potentially key moderators: paradigm, clinical status, and EEG acquisition system. It was predicted that ERN recorded during the flanker paradigm would show the highest internal consistency estimates, consistent with a previous study (Meyer et al., 2013). Considering that participants with clinical diagnoses tend to need more trials to obtain adequate internal consistency than healthy participants (e.g., Baldwin et al., 2015), it was predicted that samples including participants with clinical disorders would show poorer ERN score internal consistency than samples of healthy participants. There was no directional hypothesis regarding which acquisition systems would yield higher internal consistency scores; rather, acquisition system was included as a proxy for the various online recording characteristics that might impact internal consistency, such as active or passive electrodes. This approach was necessary, because such online recording characteristics are underreported (Clayson et al., 2019). The third aim was to examine the relationship between internal consistency and the numbers of trials retained for analysis. Lastly, exploratory analyses examined the impact of various other contextual factors (e.g., demographic characteristics and measurement approaches) on ERN score internal consistency.

2. Method

The present study hypotheses and procedures were preregistered on Open Science Framework (OSF; https://osf.io/y3jrv), and deviations from preregistered procedures are elaborated below (see Deviations from Preregistration section). The raw data and software analysis code to reproduce all analyses are also posted on OSF (https://osf.io/7jwu9/). The PRISMA guidelines for transparency and reproducibility were followed for the present meta-analysis, and the PRISMA checklist is posted on OSF.

2.1. **Literature Search and Study Selection**

The following criteria were used to include studies in this reliability generalization analysis. 1) The study examined ERN in human participants. 2) The study was written in English. 3) The study reported coefficient alpha (also known as Cronbach's alpha) estimates of recorded ERN scores, or the coefficient alpha estimates could be obtained. For example, authors of studies that examined test-retest reliability were often willing to compute coefficient alpha estimates for the purpose of this meta-analysis. 4) Internal consistency estimates were of the minimum number of trials retained for averaging or of a recommended number of trials to use as a cutoff for data inclusion/exclusion. Articles were retrieved from Web of Science, PubMed, and PsycINFO using the following search phrase: (error-related negativity OR error negativity) AND (internal consistency OR test-retest OR Cronbach's alpha OR icc OR split-half OR reliability). Searches were conducted on July 1, 2019. An additional announcement requesting internal consistency data for ERN scores was made via social media on February 19, 2020.

Additionally, the reference list of each identified article was examined for additional relevant studies. To circumvent the file-drawer problem, the corresponding authors of each article were contacted to determine whether they had any other unpublished ERN internal consistency data to contribute. Additional labs that routinely examine the internal consistency of

ERPs were also contacted to solicit unpublished ERN internal consistency data. Lastly, when any study examined test-retest reliability, had ambiguous results, or contained missing information, the corresponding author of the study was contacted for further information.

A PRISMA diagram showing the selection of studies is shown in Figure 1 (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009). A total of 106 unique articles was found from searching each database, and an additional 20 articles/datasets were received through contacting labs that routinely examine the internal consistency of ERPs, through examining the references of identified articles, or through social media. Nine articles were excluded for not reporting coefficient alpha estimates for the number of trials used as a cutoff of data inclusion (Cassidy, Robertson, & O'Connell, 2012; Chong & Meyer, 2019; DuPuis et al., 2015; Hill, Samuel, & Foti, 2016; Ip et al., 2018; Lin, 2019; Lin, Stephens, Gavin, & Davies, 2018; Riesel, Richter, Kaufmann, Kathmann, & Endrass, 2015; Segalowitz et al., 2010). Five articles were excluded for having overlapping samples with other studies (Clayson & Miller, 2017a; Larson et al., 2010; Larson, Clayson, & Baldwin, 2014a; Llerena, Wynn, Hajcak, Green, & Horan, 2016; Riesel, Weinberg, Endrass, Meyer, & Hajcak, 2013). Two articles were excluded for not examining ERN (Clayson & Larson, 2013; Marco-Pallares, Cucurell, Münte, Strien, & Rodríguez-Fornells, 2011), and two articles were excluded for being review papers (Baldwin, 2017; Clayson & Miller, 2017b). Taken together, 43 studies were included in the present meta-analysis. The author coded all studies, because the majority of studies required following up with individual authors for more information (e.g., computing coefficient alpha at trial cutoffs). All data for the present meta-analysis are posted on OSF.

<INSERT FIGURE 1 ABOUT HERE>

2.2. Data Extraction

The primary outcome measure of this study was the coefficient alpha estimate for ERN scores at the trial cutoff used for data inclusion. For example, if a study excluded all participants with fewer than eight error trials, the alpha estimate of eight ERN scores was collected for each participant sample in the study. Additionally, the number of trials used for the alpha estimates and the number of participants that were included in the estimates were also collected. Studies that reported the observed internal consistency were included (internal consistency could not be inferred based on other work).

Studies were also included when authors were willing to provide coefficient alpha estimates for the trial cutoffs that were used in either published or unpublished work. For example, some authors were willing to compute alpha estimates, despite originally inferring internal consistency based on other work. Some published studies plotted the relationship between internal consistency and the number of trials retained for averaging but did not exclude participants for having too few trials. In such instances, the number of trials needed to obtain an internal consistency estimate of .70 was used. When such information was only presented graphically, WebPlotDigitizer was used to extract internal consistency coefficients (Drevon, Fursa, & Malcolm, 2017). Lastly, studies were not excluded on any basis of potential bias/data quality, because internal consistency estimates can be considered measures of data quality. Sensitivity analyses were also used to exclude studies with highly influential estimates based on Cook's distance (Viechtbauer & Cheung, 2010).

Additional information coded for each study included 1) age, 2) sex (% female), 3) clinical status, 4) target population (i.e., undergraduate or community sample), 5) experimental paradigm (e.g., Stroop, flanker), 6) EEG reference (e.g., average reference, linked mastoid), 7) type of amplitude scoring procedure (e.g., mean amplitude, peak amplitude), 8) length of mean amplitude window (when applicable), 9) sensors used for scoring, 10) trial selection procedure for computing alpha estimates (first X number of error trials or random subset of X number of error trials), 11) whether reliability was the focal outcome of the study, 12) the approach used for ocular artifact correction, and 13) the procedure used for artifact rejection.

2.3 Internal Consistency Estimates

The number of trials retained for averaging is closely related to the observed internal consistency of ERN scores (e.g., Clayson & Miller, 2017a). Because different studies determine data inclusion based on different trial cutoffs, coefficient alpha estimates were adjusted to ensure that differences across samples were not due to the use of different numbers of trials for computing ERN score internal consistency. Hence, all reliability estimates were adjusted to the predicted coefficient alpha estimate based on eight trials using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910). The original and adjusted alpha estimates are shown for each sample in Table 1.

<INSERT TABLE 1 ABOUT HERE>

2.4 Data Analysis

An assumption of the common statistical models used in reliability generalization analyses is that effect sizes are normally distributed (Rodriguez & Maeda, 2006), and coefficient alpha estimates violate this assumption due to being bounded between 0 and 1. To circumvent the normality assumption, each alpha estimate was transformed using Bonett's transformation, which normalizes internal consistency estimates using the number of trials and participants (Bonett, 2002). All statistical models used Bonett-transformed alphas during estimation. However, for the sake of interpretability both Bonett-transformed (denoted as $\widehat{\alpha}_B$) and back-transformed estimates (denoted as $\widehat{\alpha}$) are reported for each model. When moderators were included in the model, the back-transformed estimates represent the summation of the intercept and moderator effect, again for the sake of interpretability (see Greco, O'Boyle, Cockburn, & Yuan, 2017; Piqueras, Martín-Vivar, Sandin, San Luis, & Pineda, 2017; Vicent, Rubio-Aparicio, Sánchez-Meca, & Gonzálvez, 2019).

The traditional random effects approach to meta-analysis assume that outcomes are independent from each other and only vary due to sampling variation and study variation, which results in a two-level meta-analytic model. However, an important methodological characteristic of the data used for this meta-analysis is that some studies included multiple groups of participants (e.g., a clinical group and healthy control group) or multiple alpha estimates for the same group of participants (e.g., internal consistency estimates for multiple scoring procedures or paradigms). It is likely that alpha estimates from the same study would be more similar than estimates from different studies, and treating dependent estimates as independent introduces bias by inflating the variances of the estimates, overweighting studies with multiple alpha estimates, and inflating Type I errors (Borenstein, Hedges, Higgins, & Rothstein, 2009). Hence, in order to include all internal consistency estimates without violating assumptions of statistical independence, a three-level meta-analytic model was used (Assink & Wibbelink, 2016; Cheung, 2014; Konstantopoulos, 2011). Alpha estimates for participants (Level 1), were nested with samples (Level 2), which were nested within studies (Level 3). A significant advantage of this approach is the capability to directly compare within-study and between-study moderators using all available data.

Random-effects models were used to simultaneously examine the distribution of variance across three levels and were estimated using restricted maximum likelihood (Assink & Wibbelink, 2016; Cheung, 2014). Overall variance was partitioned into variability due to sampling error (Level 1), variability due to multiple outcomes within a study (Level 2), and variability due to between-study differences (Level 3). After fitting random-effects models for Bonett-transformed alpha estimates, separate mixed-effects models were tested for each moderator. Parameters of the models were estimated using the *rma.mv* function of the *metafor* package (Viechtbauer, 2010) in *R* (R Development Core Team, 2019), and profile likelihood plots of the variance components were examined to ensure model fit. Similar to the approach for adjusting standard errors developed by Knapp and Hartung (2003), the omnibus test statistic was statistically evaluated using an *F* distribution, and moderators were statistically evaluated using a *t*-distribution (Viechtbauer, 2010). A test for residual heterogeneity without moderators in the model (Cochran's *Q* test) was used to determine whether Bonett-transformed alpha estimates were heterogeneous, and the Q_E test for residual heterogeneity for the model with moderators was used to determine whether the variability not accounted for by the moderator was larger than would be expected given the sampling variability alone (Borenstein et al., 2009; Pastor & Lazowski, 2018). When the omnibus test of moderators was significant and a moderator included more than two levels, pairwise comparisons of each level of the moderator, not including the intercept, were performed. The first level of the moderator was entered into each model as the intercept, and the *t* and *p* values presented in Tables 2 and 3 represent the test between the intercept level and the other levels.

Data Analysis Summary

In short, a three-level meta-analytic procedure was used to account for the dependencies of multiple coefficient alphas culled from single studies. Each coefficient alpha estimate was also transformed using Bonett's transformation, which normalizes internal consistency estimates using the number of trials and sample size (Bonett, 2002). For the sake of interpretability both

Bonett-transformed (denoted as $\widehat{\alpha}_B$) and back-transformed estimates (denoted as $\widehat{\alpha}$) are reported for each model. Hence, $\widehat{\alpha}$ are on the same scale as the conventional coefficient alpha and interpreted in an identical fashion.

2.5 Deviations from Preregistration

A pre-registered inclusion criterion was that coefficient alpha estimates were independent. However, many studies reported multiple internal consistency coefficients (e.g., separate coefficients for groups or measurement approaches). In order to be as inclusive as possible, all estimates from these studies were included in the meta-analysis. To include these estimates without violating independence assumptions, three-level models were used to account for the dependence of estimates obtained from the same study. Social media was also used to solicit data for the meta-analysis. Four additional moderators were coded that were not pre-registered. These additional moderators were the trial selection procedure for computing alpha estimates, whether internal consistency was the focal outcome of the study, the ocular artifact correction approach, and the procedure used for rejecting artifact.

3. Results

A total of 189 coefficient alpha estimates were culled from 68 samples nested within 43 studies. The total number of participants was 4,499 with a mean of 66 participants per study (*SD* = 100, range = 11 to 778). These data are summarized in Table 1. To ensure replicability of findings, the complete raw dataset for this reliability generalization study, including all internal consistency estimates and moderators, can be found at the OSF link provided above.

Prior to any transformation, the average of all coefficient alpha estimates was .63 (*SD* = .17, range = .02 to .91), and these estimates represent the internal consistency using an average of 10 ERN trials (*SD* = 5, range = 2 to 26; see Table 1). Given the wide variability in the numbers of trials used for estimating coefficient alpha across studies, it was *a priori* decided that estimates would be adjusted to the predicted internal consistency based on eight trials. Hence, coefficient alpha estimates were adjusted using the Spearman-Brown prophecy formula (Brown, 1910; Spearman, 1910). The mean of the coefficient alpha estimates adjusted to the predicted internal consistency of eight trials was .61 (*SD* = .17, range = .02 to .94; see Table 1).

A random-effects, intercept-only model provided an overall estimate of the average coefficient alpha as .68 (95% confidence interval [CI]: .63, .72). As predicted, the test of heterogeneity was significant, $Q(188) = 1{,}562.13$, $p < .001$. Forest plots for each internal consistency estimate are shown in Figures 2 and 3. The distribution of variance over the three levels was also examined (Cheung, 2014). Sampling error variance accounted for 9.75% of the total variance (level 1), and within-study heterogeneity was estimated as 9.49% of the total variance (level 2). Between-study heterogeneity accounted for 80.76% of the total variance (level 3). Hence, the variance not attributable to sampling error, total I_2, was 90.25%.

<INSERT FIGURES 2 & 3 ABOUT HERE>

A sensitivity analysis was conducted to determine whether any outliers might be highly influential on the overall results (Viechtbauer & Cheung, 2010). Cook's distance combines information about leverage and fit of an outcome and was used to identify outliers (Cook & Weisberg, 1982). A cutoff of $\frac{4}{k}$, where *k* is the number of outcomes (i.e., estimates) included, was used to identify influential studies. Three estimates from two studies were identified as influential in the intercept-only model (McDonald, Bozzay, Bresin, & Verona, in press; Steele et al., 2016). One estimate was likely influential due to obtaining high internal consistency with only two trials (original alpha = .81 using 2 trials; adjusted alpha = .94 using 8 trials) and a relatively large sample (*n* = 100; Steele et al., 2016). The two estimates from the other study were likely influential for having low reliability (original alphas = .11 and .44 using 2 trials; adjusted alpha = .33 and .76 using 8 trials) and a fairly large sample (n = 89; McDonald et al., in press). After removing these estimates from the intercept-only model, the estimated coefficient alpha was .67 (95% CI: .62, .71). Moving forward, models were first fit with all available data, and separate sensitivity analyses were then conducted excluding those studies with Cook's distance exceeding the specified threshold, $\frac{4}{k}$.

3.1 Publication Bias

There are no well-developed methods for detecting publication bias in three-level meta-analyses of dependent effect sizes. As a result, two methods were used to identify publication bias. First, whether a study was published was included as a moderator to determine whether published studies differed from unpublished studies. Second, it is also possible that those published studies that included internal consistency as a focal outcome would show higher internal consistency than other studies. An additional moderator analysis was conducted comparing studies wherein internal consistency was the focal outcome to those studies wherein internal consistency was not the focal outcome. Neither publication status nor focal outcome status significantly moderated internal consistency estimates, $F(1, 187) = 0.42$, $p = .52$; $F(1, 187) = 1.63$, $p = .20$, respectively (see Table 2). When influential estimates were removed, the moderator analyses for publication status and focal outcome remained nonsignificant, $F(1, 184) = 0.17$, $p = .68$; $F(1, 182) = 0.25$, $p = .62$, respectively (see Table 3).

<INSERT TABLES 2 & 3 ABOUT HERE>

3.2. Moderator Analyses

Substantial heterogeneity was observed for coefficient alpha estimates of ERN scores. Moderator analyses were conducted to identify the contextual factors that influence internal consistency, and these results are presented in Table 2. Sensitivity analyses, which removed influential estimates as outlined above, are shown in Table 3. Summaries of the number of studies, samples, and estimates and of the number of participants included in each moderator analysis are also shown in Tables 2 and 3. Notably, none of the moderators fully accounted for the observed variability in internal consistency estimates of ERN scores, as evidenced by Cochran's Q_E in Tables 2 and 3.

3.2.1 **Paradigm.** There were five different levels included in this moderator analysis. The levels included the flanker task, Go/NoGo task, picture/work task, Simon task, and Stroop task. The paradigm used for recording ERN appeared to be a significant moderator of internal consistency, $F(4, 184) = 5.77$, $p < .001$. The estimated internal consistency for the Go/NoGo task ($\widehat{\alpha} = .74$) was higher than that of the flanker task ($\widehat{\alpha} = .67$, $t(184) = 3.22$, $p = .002$), and Stroop task ($\widehat{\alpha} = .56$, $t(184) = 2.47$, $p = .01$), but it was similar to the Simon task ($\widehat{\alpha} = .73$, $t(184) = 0.14$, $p = .89$). The picture/word task level of the moderator ($\widehat{\alpha} = .27$) was significantly lower than flanker task, $t(184) = -3.01$, $p = .003$, Go/NoGo task, $t(184) = 3.71$, $p < .001$, and Simon task, $t(184) = -2.47$, $p = .01$, but it was similar to the Stroop task, $t(184) = -1.51$, $p = .13$. Significant differences in internal consistency were not observed for the Simon and Stroop tasks, $t(184) = 1.36$, $p = .18$. However, the estimate for and comparisons with the picture/word task should be interpreted cautiously due to containing only one estimate from two different samples nested within the same study. After removing 24 influential estimates from the analysis, ERN score internal consistency was no longer significantly moderated by the paradigm used for recording, $F(4, 160) = 0.53$, $p = .71$.

3.2.2 **Clinical Status.** The clinical status moderator included four levels: healthy, clinical high risk, neurological, and psychopathology. Although psychopathology groups ($\widehat{\alpha} = .60$)

demonstrated lower internal consistency than healthy groups ($\widehat{\alpha} = .69$, $t(185) = -2.04$, $p = .04$), the omnibus test of the moderator was not significant, $F(3, 185) = 1.47$, $p = .23$. Hence, initial analyses indicated that clinical status did not significantly moderate ERN score internal consistency. The sensitivity analyses yielded 20 influential estimates. After removing these influential estimates, the omnibus test of the moderator was significant, $F(3, 165) = 2.88$, $p = .04$. Internal consistency estimates from psychopathology groups ($\widehat{\alpha} = .52$) were lower than estimates from healthy groups ($\widehat{\alpha} = .67$, $t(165) = -2.77$, $p = .01$), but not significantly different from estimates from neurological groups ($\widehat{\alpha} = .68$, $t(165) = 1.97$, $p = .051$) or estimates from the clinical high risk group ($\hat{\alpha} = .76$, $t(165) = 1.76$, $p = .08$). Significant differences were not observed for the comparison between the clinical high risk group and the neurological group, $t(165) = 0.75$, $p = .45$. The estimates for the clinical high risk group are from one sample and should be interpreted with caution.

3.2.3 **EEG Acquisition System.** There were five levels of the EEG acquisition system moderator: BioSemi, ANT, Brain Products, Electrical Geodesics, Inc. (EGI), and Neuroscan. The omnibus test of moderators was not significant, $F(4, 184) = 0.57$, $p = .68$, and it remained nonsignificant after excluding 13 influential estimates from analysis, $F(4, 171) = 0.94$, $p = .45$.

3.2.4 Year of Publication. The impact of time since the publication of the first ERN score internal consistency paper was examined. The year of publication was examined as a continuous moderator and was first centered to the year 2009. Only published studies were considered in this analysis. The test of moderators was not significant, $F(1, 106) = 0.01$, $p = .92$, and this test remained nonsignificant after excluding one estimate during the sensitivity analyses, $F(1, 105) < 0.01$, $p = .98$.

3.2.5. **Age.** The mean age of the samples was 27.2 years (*SD* = 14.6, range = 9.6 to 70.8). Information for age was missing from two studies (Olvet & Hajcak, 2009; Rietdijk, Franken, & Thurik, 2014), which were excluded from this moderator analysis. Age was tested as a continuous moderator and was first centered to the minimum age of the included samples. Age did not significantly moderate ERN score internal consistency, $F(1, 185) = .004$, $p = .95$. Sensitivity analyses identified eight influential estimates. After removing these estimates from the moderator analysis, age remained nonsignificant, $F(1, 177) = 0.58$, $p = .81$.

3.2.6 **Percentage of women.** The mean percentage of women per study was 51% (*SD* = 19%, range = 0% to 100%). One study was missing information about percentage of women included (Rietdijk et al., 2014), and this study was not included in the moderator analysis. The percentage of women was tested as a continuous moderator, but it was not significant, $F(1, 186) = 1.09$, $p = .30$. Ten influential estimates were removed for the sensitivity analyses, and the percentage of women included in a sample remained nonsignificant, $F(1, 176) = 1.57$, $p = .21$.

3.2.7 **Sample type.** The sample type moderator included two levels: undergraduate sample and community sample. The omnibus test of the moderator was not significant, $F(1, 187) = 0.44$, $p = .51$. When five influential estimates were removed, the test of the moderator remained nonsignificant, $F(1, 182) = 0.22$, $p = .64$.

3.2.8 **EEG reference.** The EEG reference moderator include four levels: average reference (of all electrode sites), average ear lobes, average mastoids, and nose. The omnibus test of the moderators was significant, $F(3, 185) = 5.46$, $p = .001$, and the nose reference ($\widehat{\alpha} = .95$) showed higher estimated internal consistency than the average reference ($\widehat{\alpha} = .67$, $t(185) = 3.99$, $p < .001$), average ear lobes ($\widehat{\alpha} = .64$, $t(185) = -3.64$, $p < .001$), and average mastoids ($\widehat{\alpha} = .67$, $t(185) = -3.94$, $p < .001$). However, only one estimate was used for the nose reference level. No other contrasts were significant (|*t*s| < 0.4, *p*s > .74). The nose-reference estimate and five others were excluded during the sensitivity analysis, which subsequently yielded a nonsignificant test of moderators, *F*(2, 179) = 0.03, *p* = .97.

3.2.9 **Scoring procedure.** The amplitude scoring procedure moderator included four levels: mean, adaptive mean (mean around individual participant's peak amplitude), peak, and peak-to-peak. The test of moderators was significant, *F*(3, 185) = 12.76, *p* < .001, and each level of the moderator was significant. The estimated internal consistency ($\widehat{\alpha}$) for each level of the moderator was .67 for the mean amplitude, .70 for the adaptive mean, .71 for the peak amplitude, and .62 for the peak-to-peak amplitude. All pairwise contrasts were significant (|*t*s| > 2.4, *p*s < .02), aside from the contrast between the adaptive mean and the peak amplitude approaches, *t*(185) = -.96, *p* = .34. However, the sensitivity analyses indicated there were 45 influential estimates. Once these estimates were excluded the omnibus test of moderators was not significant, *F*(3, 140) = 1.60, *p* = .19.

3.2.10 **Length of mean.** The length of the mean amplitude window was examined as a moderator for those estimates that used either a mean or adaptive mean scoring procedure. Length of mean was tested as a continuous moderator and was first centered to the shortest mean amplitude length. The average length of the temporal window used for computing the mean amplitude was 80 ms (*SD* = 35 ms, range = 15 to 130 ms). The test of moderators was not significant, *F*(1, 156) = 1.35, *p* = .25, and this test remained nonsignificant after excluding 22 estimates for the sensitivity analyses, *F*(1, 134) < 0.01, *p* = .99.

3.2.11 **Sensors.** The sensors moderator examined whether scoring ERN amplitudes from one sensor or a cluster of sensors (i.e., region of interest [ROI]) impacted ERN score internal consistency. The omnibus test of moderators was significant, *F*(1, 187) = 5.33, *p* = .02. Single-sensor measurements of ERN scores ($\hat{\alpha}$ = .69) yielded higher internal consistency estimates than ROI measurements ($\hat{\alpha}$ = .66), t(156) =-2.31, p = .02. After excluding 26 influential estimates, and the test of moderators was not significant, F(1, 161) = 0.02, p = .90.

3.2.12 Ocular Artifact Correction. There were two levels of the ocular artifact correction moderator: independent components analysis (ICA) approaches and regression approaches. The test of moderators was significant, F(1, 187) = 5.48, p = .02. Ocular artifact correction using ICA ($\hat{\alpha}$ = .71) yielded a higher reliability estimate than correction using regression ($\hat{\alpha}$ = .57), t(187) = -2.34, p = .02. The sensitivity analyses excluded five estimates, and the test of moderators remained significant, F(1, 182) = 6.54, p = .01. Ocular artifact correction using ICA approaches similarly ($\hat{\alpha}$ = .69) resulted in higher reliability estimates than regression approaches ($\hat{\alpha}$ = .55).

3.2.13 Artifact Rejection. The approach to the rejection of artifact was also examined, and this moderator included four levels: none, automatic rejection, rejection based on visual inspection, or semiautomatic rejection (a combination of automatic rejection and rejection based on visual inspection). Although no artifact rejection ($\hat{\alpha}$ = .19) demonstrated lower reliability estimates than visual inspection ($\hat{\alpha}$ = .77) and automatic rejection ($\hat{\alpha}$ = .69, $|ts|$ > 2.0, ps < .04), the test of moderators was not significant, F(3, 185) = 1.89, p = .13. These findings should also be interpreted cautiously, because only one study did not use any artifact rejection and only two studies used only visual inspection. Eight estimates were excluded in the sensitivity analyses. Again, no artifact rejection ($\hat{\alpha}$ = .19) demonstrated lower reliability estimates than visual inspection ($\hat{\alpha}$ = .74) and automatic rejection ($\hat{\alpha}$ = .68, $|ts|$ > 2.3, ps < .03), the test of moderators was not significant, F(3, 177) = 2.28, p = .08.

3.2.14 Trial selection. The trial selection procedure refers to the approach used to estimate ERN score internal consistency. There were two approaches (i.e., levels of the moderator) examined. The first approach was scoring the first 'X' number of trials and computing internal consistency estimates for those initial trials. The second approach was to take a random subset of 'X' number of trials from all error trials and then computing an internal consistency estimate for those trials. The test of moderators was significant, $F(1, 187) = 35.97$, $p < .001$, and higher internal consistency was observed when using a subset of the beginning ERN trials ($\widehat{\alpha} = .71$) than when using a random subset of trials ($\widehat{\alpha} = .63$). The test of moderators remained significant after excluding 23 influential estimates, $F(1, 164) = 16.20$, $p < .001$. The pattern of effects for the sensitivity analyses remained the same. Internal consistency was higher when using a subset of the beginning ERN trials ($\widehat{\alpha} = .70$) than when using a random subset of trials ($\widehat{\alpha} = .62$).

3.3 Number of Trials vs. Internal Consistency

The relationship between the number of trials used for computing coefficient alpha estimates and the overall estimated alpha ($\widehat{\alpha}$) from the intercept-only random effects models was examined (see Figure 4). The top panel shows $\widehat{\alpha}$ for all included estimates. It appears that 9 and 16 trials were required to obtain an internal consistency threshold of .70 and .80, respectively. The bottom panel of Figure 4 shows the $\widehat{\alpha}$ after excluding influential estimates. The numbers of trials needed to obtain an internal consistency threshold of .70 and .80 were 10 and 17, respectively.

<INSERT FIGURE 4 ABOUT HERE>

4. Discussion

The present reliability generalization study demonstrated substantial heterogeneity in internal consistency estimates of ERN scores, and this heterogeneity was only partially accounted for by the examined moderators. There was some support for two *a priori* moderators of interest (i.e., paradigm and clinical status) and for other moderators, such as the approaches to estimating reliability and to removing ocular artifact. Using internal consistency estimates from all studies, the number of trials needed to obtain an internal consistency threshold of .80 was 16 (sensitivity analysis: 17), but this should be interpreted with caution and in the context of relevant moderators. The present findings highlight the context-dependent nature of ERN score internal consistency and the importance of evaluating internal consistency on a study-by-study basis.

Overall, the estimated coefficient alpha for eight ERN trials was .68 (95% CI: .63, .72; sensitivity analyses: .67, 95% CI: .62, .71), which is below the recommended reliability threshold of .80 for ERP research in an RDoC-type framework (Clayson & Miller, 2017b). Anecdotally speaking, six to eight trials are the most common cutoffs for inclusion of participants' data in ERN studies (see Olvet & Hajcak, 2009). Based on this meta-analysis, these cutoffs might be too low to obtain adequate internal consistency of ERN scores for most samples. Consistent with the recommendations of many others (Clayson & Miller, 2017b; Hajcak et al., 2017; Infantolino et al., 2018; Thigpen et al., 2017), the internal consistency of ERP scores needs to be calculated and reported in each study, because poor reliability can lead to mistaken inferences.

The impact of low reliability on statistical analysis can be disconcerting and dramatic. Studies using measurements with poor reliability can observe greatly *attenuated* or *exaggerated* effect sizes (i.e., magnitude error) and can observe effects that are in the opposite direction (i.e., sign error; e.g., patients > controls vs. patients < controls) from what would be observed in the population (Gelman & Carlin, 2014; Loken & Gelman, 2017; Schönbrodt & Perugini, 2013). These issues are relevant to studies of both between-group differences and within-group correlates with external variables and are especially problematic in studies with small samples (Baldwin, 2017; Brand & Bradley, 2016; Loken & Gelman, 2017; Schönbrodt & Perugini, 2013), which are common in ERP research (Clayson et al., 2019). Positive associations between internal consistency and effect sizes are observed in between-group (Hajcak et al., 2017) and within-person (Clayson & Miller, 2017a) ERN studies. Between-group effect sizes increased with increases in internal consistency in people with generalized anxiety disorder (Hajcak et al., 2017), and within-person effect sizes for correct- and error-trial ERN scores increased with increases in internal consistency (Clayson & Miller, 2017a). Taken together, using ERP data with poor score reliability can lead to mistaken statistical inferences in the form of magnitude and/or sign errors.

4.1 Moderators of ERN Score Internal Consistency

Although reliability is dependent on the population sampled, ERP score reliability is also intimately related to a host of other factors, including amplifier characteristics, recording procedures and processing pipeline, task design, and measurement approach (Boudewyn et al., 2017; Clayson, Baldwin, & Larson, 2013; Clayson & Miller, 2017b; Kappenman & Luck, 2010; Luck & Gaspelin, 2017). A difficulty that arises when considering ERN findings across studies is that each study can handle each factor differently.

The type of paradigm used to elicit ERN initially moderated internal consistency, with the Go/NoGo task ($\hat{\alpha}$ = .74) showing higher internal consistency than either the flanker task ($\hat{\alpha}$ = .67) or Stroop task ($\hat{\alpha}$ = .56) and similar internal consistency as the Simon task ($\hat{\alpha}$ = .73). This pattern of internal consistency for the Go/NoGo task, flanker task, and Stroop task is inconsistent with a previous study that showed the flanker task needed the fewest trials to achieve adequate internal consistency within the same sample of participants (Meyer et al., 2013). When only considering the task used for recording, ERN scores obtained from the same participants across these three tasks showed modest correlations, and the internal consistency across tasks varied considerably (Meyer, Bress, & Proudfit, 2014; Meyer et al., 2013; Riesel et al., 2013). When one type of paradigm is used in ERN research (e.g., a flanker task), studies can use different "flavors" of the paradigm that vary on a number of characteristics, such as the stimuli presented, timing of stimuli and response windows, number of trials, performance feedback, proportion of congruent/incongruent trials, and task instruction. This lack of standardization for recording and analyzing ERN limits the generalizability of findings across studies and remains a barrier to RDoC-inspired research (Weinberg, Dieterich, & Riesel, 2015). Different instantiations of a paradigm might lead to better or worse internal consistency, and such paradigm optimization would be a useful undertaking before making inferences about ERN score internal consistency between tasks. After removing influential estimates from the moderator analysis, the type of paradigm used no longer moderated ERN score internal consistency.

Clinical status significantly moderated ERN score internal consistency after removing influential estimates, and psychopathology groups ($\hat{\alpha}$ = .60; sensitivity analysis: $\hat{\alpha}$ = .52) showed lower internal consistency than healthy groups ($\hat{\alpha}$ = .69; sensitivity analysis: $\hat{\alpha}$ = .67). These findings suggest that psychopathology groups would generally need more trials than healthy participant groups to achieve the same level of internal consistency, which has significant implications for RDoC-inspired research. The potential cost of ignoring group differences in the psychometric properties of ERP measurements is quite high, and low score internal consistency in either a patient or control group may lead to inappropriate statistical inferences (Clayson & Miller, 2017b). A potential limitation of the present findings for ERN score internal consistency in psychopathology groups is the low representation of such research in this meta-analysis (i.e., 360 participants from 7 studies). Furthermore, the psychopathology level of the moderator represented a heterogeneous group of people with various diagnoses (see Table 1). This sparse and small representation of each diagnostic category prevented the comparison of ERN score internal consistency between psychopathology groups. The analysis on psychopathology groups as a whole, however, sheds light on the common misconception that the internal consistency of ERN scores is similar across psychopathology and healthy control groups. For example, a recent meta-analysis on the relationship between depression and ERN emphasized that not a single of the 23 examined studies evaluated the internal consistency of ERN scores (Clayson, Carbine, & Larson, 2020). Given the potential for mistaken statistical inferences in research with poor score reliability highlighted above, the adoption of new standard operating procedures that include routine evaluation of ERN score internal consistency in ERN psychopathology research seems warranted.

The lower ERN score internal consistency in psychopathology groups also leads to additional concerns above and beyond the potential for mistaken statistical inferences. Excluding patient participants based on trial or internal consistency cutoffs might bias the characteristics of the remaining sample by excluding high-performing patients (i.e., those patients with the lowest error rates and fewest ERN trials) and thereby limit generalizability. Such patients might differ on other relevant variables, such as demographic or psychiatric characteristics. Furthermore, by excluding high-performing patients, patient vs. control differences might be subsequently exaggerated by comparing only the low-performing (and possibly lower functioning) patients to the healthy controls. Unfortunately, it is virtually unknown whether excluding such high-performing patients based on trial or internal consistency cutoffs results in systematically biased samples with limited generalizability. Hence, it would be helpful if relevant characteristics of included and excluded participants were examined to determine whether using such cutoffs potentially biases the generalizability of the research.

Ocular artifact correction using ICA-based approaches ($\hat{\alpha} = .71$, sensitivity analysis: $\hat{\alpha} = .69$) yielded higher internal consistency estimates than regression-based approaches ($\hat{\alpha} = .57$, sensitivity analysis: $\hat{\alpha} = .55$). However, the two levels of this moderator represent two broad categories of ocular artifact correction. For example, there are a number of ICA-based approaches, such as rejecting ICA components based on visual inspection (Delorme & Makeig, 2004; Jung, Makeig, Bell, & Sejnowski, 1998; Jung et al., 2000), statistical criteria (Nolan, Whelan, & Reilly, 2010), or a comparison to user-defined templates (Dien, 2010), and these approaches can use different ICA algorithms to identify components. Similarly, there are various regression-based approaches to ocular artifact correction (e.g., Gratton, Coles, & Donchin, 1983; Miller, Gratton, & Yee, 1988; Semlitsch, Anderer, Schuster, & Presslich, 1986). In light of the many procedures for correcting ocular artifact, it is possible that a particular ICA- or regression-based approach might outperform others, and future research might consider identifying the best approaches in an effort to optimize the ERN data processing pipeline.

The type of EEG system used for recording ERN did not moderate internal consistency. This examination was not meant to pit one EEG system against another, but rather it was meant as a proxy for a test of online EEG recording characteristics such as type of electrodes, sampling rates, reference scheme, and filter characteristics. For example, high electrode impedance recordings are more susceptible than low impedance recordings to certain sources of noise, such as skin potentials (Kappenman & Luck, 2010). However, it is difficult to examine these online processing characteristics due to the poor reporting of ERP processing pipelines, with the typical study only reporting about two thirds of the necessary information (Clayson et al., 2019). Moving forward, when all steps of the processing pipeline are reported, it might become clearer whether particular online processing characteristics lead to improved ERN score internal consistency. Furthermore, most offline processing parameters, including EEG reference, artifact rejection approaches, ERN scoring procedure, the length of mean amplitude time window, and the sensors used for recording (ROI vs. single channel), did little to impact ERN score internal consistency, but these parameters may be more influential in the context of specific system setups (e.g., an ROI approach may yield less reliable data when using 32 channels than when using 128 channels). As such, it is again emphasized that internal consistency be presented as part of individual studies rather than assumed from previous work.

4.2 Estimation of ERN Score Internal Consistency

When internal consistency was estimated using the first subset of error trials ($\widehat{\alpha}$ = .71; sensitivity analysis: $\widehat{\alpha}$ = .70), internal consistency was higher than when it was estimated using a random subset of all error trials ($\widehat{\alpha}$ = .63; sensitivity analysis: $\widehat{\alpha}$ = .62). In a study that used multilevel modeling to look at the relationship between error trials across time and ERN amplitude, ERN decreased as participants committed more errors within a task (Volpert-Esmond, Merkle, Levsen, Ito, & Bartholow, 2017). It seems likely that ERN from the beginning error trials would be more similar in amplitude than ERN trials randomly sampled from the entire task. Hence, the actual trials selected for computing internal consistency impacts the observed estimates. This characteristic of ERN score internal consistency is similar to neuropsychological assessments, which can demonstrate substantial differences in internal consistency based on how items are used in its estimation (Kopp, Lange, & Steinke, 2019).

Although there is variability in the type of reliability coefficient used when estimating ERN score internal consistency, the present meta-analysis focused on the most widely used one, coefficient alpha. There are a number of assumptions when using coefficient alpha, such as unidimensionality, tau-equivalency, uncorrelated error variance, and an equal number of observations for each participant (Cho, 2016; Sijtsma, 2008, 2009), and these limitations for ERP score reliability estimation have been described in detail elsewhere (Baldwin et al., 2015; Clayson & Miller, 2017a, 2017b).

When it comes to estimating ERN score internal consistency, the key concern is which approach is representative of how ERN scores will be statistically analyzed. More often than not, all error trials are averaged together and then examined, which suggests that an approach that considers the pattern of responding across all trials will be more representative of the data used for statistical inferences. Taken together, internal consistency estimates from the initial errors might be overestimated due to a sampling bias. Approaches that use a single random selection might be similarly biased due to chance, and the estimates included in the meta-analysis are no exception. If a researcher would like to use classical test theory, a possible approach to circumvent sampling bias is to repeatedly randomly sample subsets of trials and examine the central tendency of the distribution of internal consistency estimates (e.g., coefficient alpha or split-half reliability) across all subsets or split halves (e.g., Larson et al., 2010). Another approach is to use all available error trials in the estimation of score internal consistency, which is possible when examining internal consistency using generalizability theory.

A significant advantage of generalizability theory over classical test theory for ERP research is the ability to use all available trials from all participants (Baldwin et al., 2015; Clayson & Miller, 2017a, 2017b), which circumvents the sampling bias endemic to coefficient alpha or split-half reliability estimates. Generalizability theory provides a multifaceted framework for examining the impact of various characteristics on internal consistency, and its goal is to pinpoint sources of systematic variability. The framework is also less restrictive than classical test theory and does not require parallel forms for estimating internal consistency, which allows for using all trials from all participants. Furthermore, generalizability theory is able to easily handle unbalanced designs, which is another common characteristic encountered in ERP studies that prevents the use of coefficient alpha estimates for all trials. The application of generalizability theory to ERP research has been explained elsewhere (Baldwin et al., 2015; Clayson & Miller, 2017a, 2017b), and the ERP Reliability Analysis (ERA) Toolbox was developed for researchers interested in applying generalizability theory approaches to ERPs (Clayson & Miller, 2017a).

4.3 Numbers of Trials

This meta-analysis estimated that 16 to 17 error trials are needed to obtain an internal consistency threshold of .80. However, these trial recommendations can be misleading, because they ignore potential moderators. For example, it is likely that ERN recorded from participants with psychopathology or studies that use regression-based approaches to correct ocular artifact will need more trials to achieve adequate internal consistency. To be clear, it is not recommended that researchers start to use these trial cutoffs for data inclusion or that these cutoffs call into question previous research with lower trial cutoffs. The number of trials retained for averaging is often inappropriately used as a proxy for internal consistency, and it is possible that some studies might demonstrate adequate internal consistency with relatively few trials. The estimates of 16 to 17 error trials should be used as guideposts when designing studies in an effort to record a sufficent number of error trials, but adequate internal consistency still needs to be verified on a study-by-study basis (Clayson & Miller, 2017b).

4.4 Limitations

There are some limitations to note. First, this reliability generalization analysis was conducted using internal consistency estimates from the minimum number of trials used for data inclusion. It is possible that once all error trials are included in analysis that internal consistency estimates for any single study would be higher. However, this assumption only holds if ERN item covariance is constant across the entire task (Cronbach, Gleser, Nanda, & Rajaratnum, 1972), which does not appear to be the case (Volpert-Esmond et al., 2017). Second, internal consistency estimates from large samples are weighted more heavily than those from small samples in the meta-analysis. Given that studies that used higher trial cutoffs for data inclusion likely excluded more participants, it is possible that the meta-analytic internal consistency estimates might be biased upward by the exclusion of participants with poorer score internal consistency. Third, with regard to the sensors moderator, some studies used large ROIs that covered a large portion of the scalp. ERN score internal consistency should improve only insofar as the signal of interest is being captured by the ROI (Clayson & Miller, 2017b), and the inclusion of studies with large ROIs or few electrodes spaced far apart might have led to better internal consistency estimates for a single sensor than for an ROI. Fourth, some levels of some moderators had very few estimates included in the analysis, and such findings should be interpreted with caution. Similarly, there is overlap in levels between some moderators that prevent interaction effects from being meaningfully examined. As more studies begin to routinely report ERN score internal consistency, all levels of these moderators and interactions between moderators could be analyzed in future reliability generalization studies.

Lastly, some research labs are interested in the psychometric analysis of ERN and have published multiple such studies. Other labs have a routine practice of reporting internal consistency estimates of ERPs. As a result, internal consistency estimates from these labs constitute a large portion of data in the present meta-analysis. When focal outcome, whether the focus of the study was the internal consistency of ERN scores, was examined as a moderator, the initial analysis and sensitivity analysis were not significant. This suggests that psychometric studies of ERN scores are likely not inflated due to publication bias. However, it is a possibility that such labs that conduct psychometric studies and routinely report internal consistency estimates give greater attention to the data processing steps that yield better reliability estimates. However, until internal consistency is routinely reported, and possibly until EEG data become more widely shared, it is difficult to know whether the internal consistency estimates included in this meta-analysis are inflated due to a reporting bias.

4.5 Moving Forward

Although inferring reliability based on previous psychometric research is a widespread practice, the substantial heterogeneity of ERN score internal consistency calls this practice into question. Contextual factors are clearly important. Furthermore, this practice is inappropriate on theoretical grounds, because score reliability is the property of the data in hand, not the property of a particular measure or ERP (Thompson, 2003; Vacha-Haase, 1998). Poor internal consistency substantially limits the utility of ERPs as dimensional measures for RDoC-inspired research, because the internal consistency of measurements is closely related to how well measurements can differentiate among participants. Hence, measures with poor internal consistency are poorly suited for studying individual differences.

Simply including more trials is an unlikely panacea for problematic ERN score internal consistency, because the relationship between internal consistency and the number of trials included in subject averages asymptotes (Clayson & Miller, 2017a). Adequate internal consistency can be achieved with few trials when data have a high signal-to-noise ratio. For example, there was a wide range of internal consistency estimates in the studies examined (see Table 1), and some studies were able to achieve adequate internal consistency with fewer trials than others. Thus, any efforts toward improving the signal-to-noise ratio of EEG data during recording, processing, and analysis should benefit score reliability (Boudewyn et al., 2017; Clayson & Miller, 2017b; Kappenman & Luck, 2010; Luck & Gaspelin, 2017; Thigpen et al., 2017). The present meta-analysis provides support for using ICA-based ocular artifact correction over regression-based ocular artifact correction for improving ERN score internal consistency. Future research that examines the impact of data recording, processing, and analysis procedures on ERN score reliability would be helpful for optimizing paradigms for the study of individual differences (e.g., Klawohn, Meyer, Weinberg, & Hajcak, 2020; Sandre, Banica, Riesel, Klawohn, & Weinberg, under review), and such research could consider any of the data processing steps outlined in the recent ERP publication guidelines paper (Keil et al., 2014) as potential moderators of internal consistency.

The approach to estimating ERN score internal consistency substantially impacted observed estimates. Moving forward approaches to estimating internal consistency that minimize the potential for sampling error should be used. If a researcher is interested in using classical test theory approaches, a coefficient alpha could be computed on numerous random samples of ‘X’ number of ERN scores without replacement from each participant, and the central tendency of estimates could then be used to identify appropriate cutoffs (see Larson et al., 2010). When using a split-half reliability estimates, numerous different split halves (e.g., odd vs even trials or first half vs. second half) could be computed in the same fashion, because any one split-half estimate is still suspect due to sampling bias. This is particularly the case when few trials are used in each split half, which is common for ERN studies. Another approach is to use generalizability theory estimates of internal consistency, and the advantage of this approach is that these estimates use all trials from all participants, which circumvents the sampling bias endemic to classical test theory approaches (Baldwin et al., 2015; Clayson & Miller, 2017a, 2017b). The ERA Toolbox is open-source MATLAB software that can compute internal consistency of ERP scores using generalizability theory, and it was specifically developed with ERP scores in mind (Clayson & Miller, 2017a).

Although the framing of this reliability generalization study focused on individual differences (i.e., RDoC-inspired research), it is also important to ensure that internal consistency is similar across groups, when between-group differences are of interest. Group differences in ERN score internal consistency have been observed between healthy and psychopathology groups in all published psychometric evaluations of ERN (e.g., Baldwin et al., 2015; Foti et al., 2013; Hajcak et al., 2017), and this meta-analysis confirmed such differences. Ignoring between-group differences in the internal consistency of ERP scores can lead to mistaken statistical inferences (Brakenhoff, van Smeden, Visseren, & Groenwold, 2018; Gelman & Carlin, 2014; Loken & Gelman, 2017). For example, group differences can be observed simply due to poor score reliability in one or both groups (Cooper et al., 2017; Hedge et al., 2017). Thus, it is important to ensure similar score reliability across groups.

Anecdotally speaking, a barrier to reporting internal consistency estimates that was apparent after contacting many authors for data was the inability of popular EEG/ERP processing software to easily compute reliability estimates. Computing internal consistency estimates typically requires exporting single-trial estimates for each event and person, compiling the estimates into a single dataset, and using statistical software packages to calculate the estimates. These first two steps can be a substantial obstacle to overcome using some software, unless the user has programming or data management expertise. The reporting of reliability estimates in the literature would likely improve if software developers facilitated the exporting of compiled single-trial estimates (in a wide format, separate single-trial measurements down rows, and in a long format, separate single-trial measurements across columns) or included functions to compute reliability estimates.

The substantial heterogeneity in ERN score internal consistency estimates definitively demonstrates that internal consistency cannot be inferred by obtaining a trial threshold recommended from a previous psychometric analysis. Unfortunately, this practice is widespread. The present analyses of 4,499 participants from 43 studies indicates that at least 16-17 trials are needed to obtain a coefficient alpha of .80, and these numbers are much higher than the commonly used thresholds of six to eight trials based off of studies with many fewer participants than this meta-analysis (see Table 1). The implications of these findings for previous ERN research on individual differences is disconcerting. However, some studies are able to obtain adequate internal consistency with fewer than eight trials (e.g., Pontifex et al., 2010; Seer et al., 2017; Steele et al., 2016). As such, the present analyses should not be used to oppugn prior research that used low trial cutoffs, because doing so would be based on the same fallacious practice of inferring reliability based on trial cutoffs. The glaring issue in the literature is the failure to routinely report internal consistency estimates in ERN studies of individual differences. That is a practice that must change. Notably, the outlook moving forward appears hopeful, because journals are beginning to adopt guidelines for the reporting of internal consistency on a study-by-study basis.

Hence, the last recommendation is to routinely report ERP score reliability in all research, particularly individual differences research. Considering the importance of measurement internal consistency in RDoC-inspired research of dimensional constructs, the routine reporting of internal consistency is valuable for identifying candidate measures. Furthermore, the considerable heterogeneity in ERN score internal consistency supports the adoption and enforcement of guidelines for routinely reporting score reliability of psychophysiological measures (e.g., author guidelines of *Psychophysiology* and *International Journal of Psychophysiology*). Reliability cannot be inferred based on previous psychometric analyses. Within a single study, information about internal consistency provides a context for interpreting statistical inferences (LeBel & Paunonen, 2011; Thompson, 2003; Wilkinson & The APA Task Force on Statistical Inference, 1999). Across studies, such information aids in the selection of ERP components and paradigms for examining individual differences and can be synthesized in reliability generalization studies to determine moderators of internal consistency.

Acknowledgments

I would like to thank the many authors that went above and beyond to provide me with clarifying information and additional data for this reliability generalization study. I would also like to thank Michael J. Larson, Ph.D., and Scott A. Baldwin, Ph.D., for comments on an earlier draft of the manuscript. Writing of this manuscript was supported in part by the Office of Academic Affiliations, Advanced Fellowship Program in Mental Illness Research and Treatment, Department of Veterans Affairs.

References

Assink, M., & Wibbelink, C. J. M. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *The Quantitative Methods for Psychology, 12*, 154-174. doi:10.20982/tqmp.12.3.p154

*Bailey, B., & Larson, M. J. (in prep). *The impact of intense exercise on ERN ampltiude*.

Baldwin, S. A. (2017). Improving the rigor of psychophysiology research. *Int J Psychophysiol, 111*, 5-16. doi:10.1016/j.ijpsycho.2016.04.006

*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*, 790-800. doi:10.1111/psyp.12401

Bonett, D. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*, 335-340. doi:10.3102/10769986027004335

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley.

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15*, 386-397. doi:10.1037/a0019626

*Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2017). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology, 14*, e13049. doi:10.1111/psyp.13049

Brakenhoff, T. B., van Smeden, M., Visseren, F. L. J., & Groenwold, R. H. H. (2018). Random measurement error: Why worry? An example of cardiovascular risk factors. *PloS one, 13*, e0192298-0192298. doi:10.1371/journal.pone.0192298

Brand, A., & Bradley, M. T. (2016). The precision of effect size estimation from published psychological research: Surveying confidence intervals. *Psychological Reports, 118*, 154-170. doi:10.1177/0033294115625265

*Bresin, K., & Verona, E. (in press). Craving and substance use: Examining psychophysiological and behavioral moderators. *International Journal of Psychophysiology*. doi:10.1016/j.ijpsycho.2019.03.006

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*, 196-322. doi:10.1111/j.2044-8295.1910.tb00207.x

*Burwell, S. J., Malone, S. M., & Iacono, W. G. (2016). One-year developmental stability and covariance among oddball, novelty, go/no-go, and flanker event-related potentials in adolescence: A monozygotic twin study. *Psychophysiology, 53*, 991-1007. doi:10.1111/psyp.12646

*Carbine, K. A., & Larson, M. J. (in prep). *ERN recorded during a Go/NoGo task with food stimuli*.

Cassidy, S. M., Robertson, I. H., & O'Connell, R. G. (2012). Retest reliability of event-related potentials: Evidence from a variety of paradigms. *Psychophysiology, 49*, 659-664. doi:10.1111/j.1469-8986.2011.01349.x

*Cavanagh, J. F., Gründler, T. O. J., Frank, M. J., & Allen, J. J. B. (2010). Altered cingulate sub-region activation accounts for task-related dissociation in ERN amplitude as a function of obsessive-compulsive symptoms. *Neuropsychologia, 48*, 2098-2109.

*Cavanagh, J. F., Masters, S. E., Bath, K., & Frank, M. J. (2014). Conflict acts as an implicit cost in reinforcement learning. *Nature Communications, 5*, 1-10.

*Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. B. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology, 49*, 220-238. doi:10.1111/j.1469-8986.2011.01293.x

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods, 19*, 211-229. doi:10.1037/a0032968

Cho, E. (2016). Making reliability reliable. *Organizational Research Methods, 19*, 651 - 682. doi:10.1177/1094428116656239

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

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*, 174-186. doi:10.1111/psyp.12001

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, 111*, 5-17. doi:10.1111/psyp.13437

Clayson, P. E., Carbine, K. A., & Larson, M. J. (2020). Error-related negativity and reward positivity as biomarkers of depression: P-curving the evidence. *International Journal of Psychophysiology, 150*, 50-72. doi:10.1016/j.ijpsycho.2020.01.005

Clayson, P. E., & Larson, M. J. (2013). Psychometric properties of conflict monitoring and conflict adaptation indices: Response time and conflict N2 event-related potentials. *Psychophysiology, 50*, 1209-1219. doi:10.1111/psyp.12138

*Clayson, P. E., & Larson, M. J. (2019). The impact of recent and concurrent affective context on cognitive control: An ERP study of performance monitoring. *International Journal of Psychophysiology, 143*, 44-56. doi:10.1016/j.ijpsycho.2019.06.007

*Clayson, P. E., & Larson, M. J. (in prep). *ERN across psychopathology groups*.

Clayson, P. E., & Miller, G. A. (2017a). ERP Reliability Analysis (ERA) Toolbox: An open-source toolbox for analyzing the reliability of event-related potentials. *International Journal of Psychophysiology, 111*, 68-79. doi: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. doi:10.1016/j.ijpsycho.2016.09.005

*Clayson, P. E., Wynn, J. K., Green, M. F., & Horan, W. P. (2018). Cognitive correlates of performance monitoring ERPs in Schizophrenia. *Psychophysiology, 55*, S34-S136. doi:10.1111/psyp.13264

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*: New York: Chapman and Hall.

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The role of psychometrics in individual differences research in cognition: A case study of the AX-CPT. *Frontiers in Psychology, 8*, 136-116. doi:10.3389/fpsyg.2017.01482

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnum, N. (1972). *The dependability of behavioral measures: Theory of generalizability for scores and profiles*. New York, NY: John Wiley.

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine, 11*, 768. doi:10.1186/1741-7015-11-126

Cuthbert, B. N., & Kozak, M. J. (2013). Constructing constructs for psychopathology: The NIMH research domain criteria. *Journal of Abnormal Psychology, 122*, 928-937. doi:10.1037/a0034028

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analsis. *Journal of Neuroscience Methods, 134*, 9-21.

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*, 138-145. doi:10.1016/j.jneumeth.2009.12.009

Drevon, D., Fursa, S. R., & Malcolm, A. L. (2017). Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data. *Behavior Modification, 41*, 323-339. doi:10.1177/0145445516673998

DuPuis, D., Ram, N., Willner, C. J., Karalunas, S., Segalowitz, S. J., & Gatzke-Kopp, L. M. (2015). Implications of ongoing neural development for the measurement of the error-related negativity in childhood. *Developmental Science, 18*, 452-468. doi:10.1111/desc.12229

*Elkins-Brown, N., Saunders, B., & Inzlicht, M. (2018). The misattribution of emotions and the error-related negativity: A registered report. *Cortex, 109*, 124-140. doi:10.1016/j.cortex.2018.08.017

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Banke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology, 78*, 447-455.

*Fischer, A. G., Klein, T. A., & Ullsperger, M. (2017). Comparing the error-related negativity across groups: The impact of error- and trial-number differences. *Psychophysiology, 54*, 998-1009. doi:10.1111/psyp.12863

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences of the United States of America, 127*, 201711978-201711910. doi:10.1073/pnas.1711978115

Flegal, K. M., Kit, B. K., & Graubard, B. I. (2017). Bias in hazard ratios arising From misclassification according to self-reported weight and height in observational studies of body mass index and mortality. *American Journal of Epidemiology, 187*, 125-134. doi:10.1093/aje/kwx193

*Foti, D., Kotov, R., & Hajcak, G. (2013). Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. *Journal of Abnormal Psychology, 122*, 520-531. doi:10.1037/a0032618

Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability fallacy: Similar group effects may arise from unreliable individual effects. *NeuroImage, 195*, 174-189.

*García Alanis, J. C., Baker, T. E., Peper, M., & Chavanon, M.-L. (2019). Social context effects on error-related brain activity are dependent on interpersonal and achievement-related traits. *Scientific Reports, 9*, 1728. doi:10.1038/s41598-018-38417-2

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641-651. doi:10.1177/1745691614551642

*Glazer, J. E., & Nusslock, R. (unpublished). *Brain, motivation, and personality development research project*.

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

Greco, L. M., O'Boyle, E. H., Cockburn, B. S., & Yuan, Z. (2017). Meta-analysis of coefficient alpha: A reliability generalization study. *Journal of Management Studies, 55*, 583-618. doi:10.1111/joms.12328

*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*, 823-834. doi:10.1037/abn0000274

Hedge, C., Powell, G., & Sumner, P. (2017). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 103*, 411-421. doi:papers3://publication/doi/10.3758/s13428-017-0935-1

Hill, K. E., Samuel, D. B., & Foti, D. (2016). Contextualizing individual differences in error monitoring: Links with impulsivity, negative affect, and conscientiousness. *Psychophysiology, 53*, 1143-1153. doi:10.1111/psyp.12671

Infantolino, Z. P., Luking, K. R., Sauder, C. L., Curtin, J. J., & Hajcak, G. (2018). Robust is not necessarily reliable: From within-subjects fMRI contrasts to between-subjects comparisons. *NeuroImage, 173*, 146-152. doi:10.1016/j.neuroimage.2018.02.024

Ip, C.-T., Ganz, M., Ozenne, B., Sluth, L. B., Gram, M., Viardot, G., . . . Christensen, S. R. (2018). Pre-intervention test-retest reliability of EEG and ERP over four recording intervals. *International Journal of Psychophysiology, 134*, 30-43. doi:10.1016/j.ijpsycho.2018.09.007

Jung, T.-P., Makeig, S., Bell, A. J., & Sejnowski, T. J. (1998). Independent component anlysis of electroencephalographic and event-related potential data. In P. Poon & J. Brugge (Eds.), *Auditory processing and neural modeling* (pp. 189-197). New York: Plenum Press.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology, 37*, 163-178.

Kappenman, E. S., & Luck, S. J. (2010). The effects of electrode impedance on data quality and statistical significance in ERP recordings. *Psychophysiology, 47*, 888-904. doi:10.1111/j.1469-8986.2010.01009.x

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., . . . Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology, 51*, 1-21.

Klawohn, J., Meyer, A., Weinberg, A., & Hajcak, G. (2020). Methodological choices in event-related potential (ERP) research and their impact on internal consistency reliability and individual differences: An examination of the error-related negativity (ERN) and anxiety. *Journal of Abnormal Psychology, 129*, 29-37. doi:10.1037/abn0000458

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine, 22*, 2693-2710. doi:10.1002/sim.1482

Kolossa, A., & Kopp, B. (2018). Data quality over data quantity in computational cognitive neuroscience. *NeuroImage, 172*, 775-785. doi:10.1016/j.neuroimage.2018.01.005

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2*, 61-76. doi:10.1002/jrsm.35

Kopp, B., Lange, F., & Steinke, A. (2019). The reliability of the Wisconsin Card Sorting Test in clinical practice. *Assessment, 35*, 1-16. doi:10.1177/1073191119866257

Kozak, M. J., & Cuthbert, B. N. (2016). The NIMH Research Domain Criteria Initiative: Background, issues, and pragmatics. *Psychophysiology, 53*, 286-297. doi:10.1111/psyp.12518

*Larson, M. J. (unpublished-a). *ERN during pregnancy*.

*Larson, M. J. (unpublished-b). *ERN during the Stroop task*.

*Larson, M. J. (unpublished-c). *ERN in children and adolescents*.

*Larson, M. J. (unpublished-d). *ERN in older adults*.

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

*Larson, M. J., & Clayson, P. E. (in prep). ERN in mild traumatic brain injury.

Larson, M. J., Clayson, P. E., & Baldwin, S. A. (2014a). How dependable are electrophysiological indices of performance monitoring (ERN, Pe) in a clinical sample? A generalizability and decision theory analysis. *Psychophysiology, 51*, S14-S79. doi:10.1111/psyp.12280

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

*Larson, M. J., Clayson, P. E., & Farrer, T. J. (2012). Performance monitoring and cognitive control in individuals with mild traumatic brain injury. *Journal of the International Neuropsychological Society, 18*, 323-333. doi:10.1017/S1355617711001779

*Larson, M. J., Perry, C. E., Hedges, D. W., Nielsen, B. L., Holt-Lunstad, J., & Call, V. R. A. (2014c). The interaction between dopamine D2 receptor gene, telomere length, and ERP indices of performance monitoring in community-dwelling older adults. *Psychophysiology, 51*, S14-S79. doi:10.1111/psyp.12280

LeBel, E. P., & Paunonen, S. V. (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin, 37*, 570-583. doi:10.1177/0146167211400619

Lin, M.-H. (2019). *The relationship between cognitive functions and occupational performance in children, adults, and adults with Attention Deficit Hyperactivity Disorders (ADHD).*

Lin, M.-H., Stephens, J., Gavin, W., & Davies, P. (2018). The test-retest reliability of the error-related negativity (ERN) and error positivity (Pe) amplitudes in neurotypical children and adults. *Psychophysiology, 55*, S34-S136. doi:10.1111/psyp.13264

Llerena, K., Wynn, J. K., Hajcak, G., Green, M. F., & Horan, W. P. (2016). Patterns and reliability of EEG during error monitoring for internal versus external feedback in schizophrenia. *International Journal of Psychophysiology, 105*, 39-46. doi:10.1016/j.ijpsycho.2016.04.012

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

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

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

*McDonald, J. B., Bozzay, M. L., Bresin, K., & Verona, E. (in press). Facets of externalizing psychopathology in relation to inhibitory control and error processing. *International Journal of Psychophysiology*. doi:10.1016/j.ijpsycho.2019.08.007

Meyer, A., Bress, J. N., & Proudfit, G. H. (2014). Psychometric properties of the error-related negativity in children and adolescents. *Psychophysiology, 51*, 602-610. doi:10.1111/psyp.12208

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

Miller, G. A., Gratton, G., & Yee, C. M. (1988). Generalized implementation of an eye movement correction procedure. *Psychophysiology, 25*, 241-243. doi:10.1111/j.1469-8986.1988.tb00999.x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Journal of Clinical Epidemiology, 62*, 1006-1012. doi:10.1016/j.jclinepi.2009.06.005

*Moser, J. (in prep). *ERN in those at clinical high risk for developing psychopathology*.

*Moser, J., Clayson, P. E., Muir, A., Allen, W., Modersitzki, E., Louis, C., . . . Larson, M. J. (2019). A preliminary report on a multi-site effort to replicate the association between worry and enlarged actionam monitoring brain potentials (Moser et al., 2012). *Psychophysiology, 56*, S3-S39. doi:10.1111/psyp.13501

*Muir, A. M., Carbine, K. A., Goodwin, J., Hedges-Muncy, A., Endrass, T., & Larson, M. J. (2019). Differentiating electrophysiological indices of internal and external performance monitoring: Relationship with perfectionism and locus of control. *PloS one, 14*, e0219883-0219821. doi:10.1371/journal.pone.0219883

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology, 38*, 752-760. doi:10.1111/1469-8986.3850752

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully automated statistical thresholding for EEG artifact rejection. *Journal of Neuroscience Methods, 192*, 152 - 162. doi:10.1016/j.jneumeth.2010.07.015

*Olson, R. L., Brush, C. J., Ehmann, P. J., Buckman, J. F., & Alderman, B. L. (2018). A history of sport-related concussion is associated with sustained deficits in conflict and error monitoring. *International Journal of Psychophysiology, 132*, 145-154. doi:10.1016/j.ijpsycho.2018.01.006

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

Pastor, D. A., & Lazowski, R. A. (2018). On the multilevel nature of meta-analysis: A tutorial, comparison of software programs, and discussion of analytic choices. *Multivariate Behavioral Research, 53*, 74-89. doi:10.1080/00273171.2017.1365684

Piqueras, J. A., Martín-Vivar, M., Sandin, B., San Luis, C., & Pineda, D. (2017). The Revised Child Anxiety and Depression Scale: A systematic review and reliability generalization meta-analysis. *Journal of Affective Disorders, 218*, 153-169. doi:10.1016/j.jad.2017.04.022

*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*, 767-773. doi:10.1111/j.1469-8986.2010.00974.x

R Development Core Team. (2019). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Riesel, A., Richter, A., Kaufmann, C., Kathmann, N., & Endrass, T. (2015). Performance monitoring in obsessive-compulsive undergraduates: Effects of task difficulty. *Brain and Cognition, 98*, 35-42. doi:10.3758/s13423-018-1558-y

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

*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*, e102672-102677. doi:10.1371/journal.pone.0102672

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322. doi:10.1037/1082-989X.11.3.306

Rouder, J., & Haaf, J. M. (2018). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review, 26*, 452-467. doi:10.31234/osf.io/f3h2k

*Sandre, A., Banica, I., Riesel, A., Klawohn, J., & Weinberg, A. (under review). Comparing the effects of different methodological decisions on the error-related negativity and its association with behavior and gender.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609-612. doi:10.1016/j.jrp.2013.05.009

*Seer, C., Lange, F., Loens, S., Wegner, F., Schrader, C., Dressler, D., . . . Kopp, B. (2017). Dopaminergic modulation of performance monitoring in Parkinson’s disease: An event-related potential study. *Nature Publishing Group*, 1-13. doi:10.1038/srep41222

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

Seghier, M. L., & Price, C. J. (2018). Interpreting and utilising intersubject variability in brain function. *Trends in Cognitive Sciences, 22*, 517-530. doi:10.1016/j.tics.2018.03.003

Semlitsch, H. V., Anderer, P., Schuster, P., & Presslich, O. (1986). A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. *Psychophysiology, 23*, 695 - 703. doi:10.1111/j.1469-8986.1986.tb00696.x

Sijtsma, K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach’s alpha. *Psychometrika, 74*, 107-120. doi:10.1007/s11336-008-9101-0

Sijtsma, K. (2009). Reliability beyond theory and into practice. *Psychometrika, 74*, 169-173. doi:10.1007/s11336-008-9103-y

*Singh, A., Richardson, S. P., Narayanan, N., & Cavanagh, J. F. (2018). Mid-frontal theta activity is diminished during cognitive control in Parkinson's disease. *Neuropsychologia, 117*, 113-122. doi:10.1016/j.neuropsychologia.2018.05.020

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295. doi:10.1111/j.2044-8295.1910.tb00206.x

*Steele, V. R., Anderson, N. E., Claus, E. D., Bernat, E. M., Rao, V., Assaf, M., . . . Kiehl, K. A. (2016). Neuroimaging measures of error-processing: Extracting reliable signals from event-related potentials and functional magnetic resonance imaging. *NeuroImage, 132*, 247-260. doi:10.1016/j.neuroimage.2016.02.046

*Suchan, F., Kopf, J., Althen, H., Reif, A., & Plichta, M. M. (2018). Reliable and efficient recording of the error-related negativity with a speeded Eriksen Flanker Task. *Acta Neuropsychiatrica, 31*, 1-8. doi:10.1017/neu.2018.36

Thigpen, N. N., Kappenman, E. S., & Keil, A. (2017). Assessing the internal consistency of the event-related potential: An example analysis. *Psychophysiology, 54*, 123-138. doi:10.1111/psyp.12629

Thompson, B. (2003). Guidelines for authors reporting score reliability estimates. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 91-101). Thousand Oaks, CA: Sage Publications, Inc.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*, 6-20. doi:10.1177/0013164498058001002

*Valadez, E. A., & Simons, R. F. (2018). The power of frontal midline theta and post-error slowing to predict performance recovery: Evidence for compensatory mechanisms. *Psychophysiology, 55*, e13010. doi:10.1111/psyp.13010

Vicent, M., Rubio-Aparicio, M., Sánchez-Meca, J., & Gonzálvez, C. (2019). A reliability generalization meta-analysis of the child and adolescent perfectionism scale. *Journal of Affective Disorders, 245*, 533-544. doi:10.1016/j.jad.2018.11.049

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112-125. doi:10.1002/jrsm.11

Volpert-Esmond, H. I., Merkle, E. C., Levsen, M. P., Ito, T. A., & Bartholow, B. D. (2017). Using trial-level data and multilevel modeling to investigate within-task change in event-related potentials. *Psychophysiology, 55*, e13044-13012. doi:10.1111/psyp.13044

*Warren, C., Seer, C., Lange, F., Kopp, B., & Müller-Vahl, K. (2020). Neural correlates of performance monitoring in adult patients with Gilles de la Tourette syndrome: A study of event-related potentials. *Clinical Neurophysiology, 131*, 597-608.

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

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and expectations. *American Psychologist, 54*, 594-604. doi:10.1037/0003-066X.54.8.594

*Xu, X., & Inzlicht, M. (2015). Neurophysiological responses to gun-shooting errors. *International Journal of Psychophysiology, 95*, 247-253. doi:10.1016/j.ijpsycho.2014.10.015

Table 1

Summary Information for Included Datasets

Author	Study Number	*n*	Mean Age	% Female	Diagnosis	Group	Sample Type	# of Estimates	Original Reliability	SB Reliability	Trials
Bailey and Larson (in prep)	1	144	22.14	54%	--	healthy	undergraduate	9	0.69 (0.44, 0.82)	0.65 (0.36, 0.77)	10 (8, 11)
Baldwin et al. (2015)	2	239	21.93	52%	--	healthy	community	2	0.78 (0.76, 0.80)	0.59 (0.56, 0.62)	20 (20, 20)
		31	21.87	65%	MDD	psychopathology	community	2	0.66 (0.52, 0.79)	0.45 (0.30, 0.60)	20 (20, 20)
		23	22.61	83%	Anxiety Disorders	psychopathology	community	2	0.52 (0.52, 0.52)	0.30 (0.30, 0.30)	20 (20, 20)
Boudewyn et al. (2017)	3	32	21	66%	--	healthy	undergraduate	1	0.70	0.70	8
Bresin and Verona (in press)	4	43	30.09	53%	SUDs/AUDs	psychopathology	community	2	0.28 (0.27, 0.29)	0.61 (0.6, 0.62)	2 (2, 2)
		11	21.63	55%	--	healthy	community	1	0.06	0.2	2
Burwell et al. (2016)	5	85	15.4	51%	--	healthy	community	2	0.56 (0.55, 0.58)	0.63 (0.62, 0.65)	6 (6, 6)
Carbine and Larson (in prep)	6	48	19.65	65%	--	healthy	undergraduate	3	0.71 (0.67, 0.73)	0.80 (0.76, 0.81)	5 (5, 5)
Cavanagh et al. (2010)	7	23	19	48%	--	healthy	undergraduate	4	0.16 (0.07, 0.32)	0.15 (0.06, 0.29)	9 (9, 9)
		23	19.13	30%	OCD	psychopathology	undergraduate	4	0.24 (0.02, 0.34)	0.22 (0.02, 0.31)	9 (9, 9)
Cavanagh et al. (2012)	8	40	19.18	30%	--	healthy	undergraduate	4	0.49 (0.29, 0.76)	0.77 (0.62, 0.93)	2 (2, 2)
Cavanagh et al. (2014)	9	67	20.26	36%	--	healthy	undergraduate	2	0.31 (0.27, 0.35)	0.64 (0.6, 0.68)	2 (2, 2)
		32	20.91	44%	--	healthy	undergraduate	4	0.48 (0.32, 0.68)	0.78 (0.65, 0.89)	2 (2, 2)
Clayson et al. (2018)	10	52	47.79	40%	--	healthy	community	2	0.64 (0.61, 0.67)	0.67 (0.64, 0.70)	7 (7, 7)
		60	50.17	22%	Schizophrenia	psychopathology	community	2	0.69 (0.67, 0.71)	0.54 (0.52, 0.57)	15 (15, 15)
Clayson and Larson (2019)	11	28	21	50%	--	healthy	undergraduate	2	0.72 (0.71, 0.73)	0.56 (0.55, 0.57)	16 (16, 16)

		31	21	62%	--	healthy	undergraduate	2	0.72 (0.61, 0.83)	0.62 (0.49, 0.75)	13 (13, 13)
Clayson and Larson (in prep)	12	21	22.71	52%	OCD	psychopathology	community	2	0.68 (0.66, 0.69)	0.45 (0.44, 0.47)	20 (20, 20)
		29	21.9	80%	MDD	psychopathology	community	2	0.62 (0.56, 0.67)	0.40 (0.35, 0.46)	19 (19, 19)
		29	21.66	93%	GAD	psychopathology	community	2	0.57 (0.53, 0.61)	0.37 (0.33, 0.41)	18 (18, 18)
		27	21.04	62%	--	healthy	undergraduate	2	0.77 (0.71, 0.84)	0.65 (0.57, 0.74)	15 (15, 15)
Elkins-Brown et al. (2018)	13	38	20.02	58%	--	healthy	undergraduate	2	0.68 (0.68, 0.69)	0.74 (0.74, 0.75)	6 (6, 6)
		39	19.97	69%	--	healthy	undergraduate	2	0.55 (0.45, 0.64)	0.61 (0.52, 0.70)	6 (6, 6)
Fischer et al. (2017)	14	778	24.1	50%	--	healthy	community	6	0.90 (0.89, 0.91)	0.82 (0.78, 0.85)	16 (14, 18)
Foti et al. (2013)	15	76	43.34	33%	Psychotic Disorders	psychopathology	community	2	0.68 (0.68, 0.69)	0.62 (0.46, 0.78)	12.5 (5, 20)
		52	39	50%	--	healthy	community	2	0.68 (0.65, 0.72)	0.57 (0.43, 0.72)	14 (8, 20)
García Alanis et al. (2019)	16	30	24	0%	--	healthy	undergraduate	1	0.70	0.76	6
		35	23	0%	--	healthy	undergraduate	1	0.68	0.74	6
Glazer and Nusslock (unpublished)	17	54	20.24	70%	--	healthy	community	2	0.67 (0.66, 0.67)	0.67 (0.66, 0.67)	8 (8, 8)
Hajcak et al. (2017)	18	36	23.58	83%	--	healthy	community	1	0.70	0.47	21
		25	26.48	96%	GAD	psychopathology	community	1	0.70	0.42	26
Larson et al. (2012)	19	33	21.84	42%	mTBI	neurological	community	4	0.58 (0.50, 0.68)	0.44 (0.36, 0.55)	14 (14, 14)
		44	20.77	52%	--	healthy	community	4	0.72 (0.70, 0.75)	0.60 (0.57, 0.63)	14 (14, 14)
Larson et al. (2014c)	20	90	21.78	46%	--	healthy	community	8	0.61 (0.49, 0.72)	0.61 (0.49, 0.72)	8 (8, 8)
Larson and Clayson (in prep)	21	48	19.92	48%	--	healthy	community	8	0.66 (0.52, 0.72)	0.50 (0.35, 0.56)	16 (16, 16)
		59	20.46	61%	mTBI	neurological	community	8	0.64 (0.37, 0.76)	0.62 (0.34, 0.74)	9 (9, 9)

Larson (unpublished-a)	22	29	23.34	100%	--	healthy	undergraduate	4	0.64 (0.57, 0.67)	0.64 (0.57, 0.67)	8 (8, 8)
Larson (unpublished-b)	23	104	20.37	49%	--	healthy	undergraduate	4	0.58 (0.50, 0.65)	0.58 (0.5, 0.65)	8 (8, 8)
Larson (unpublished-c)	24	41	11.76	66%	--	healthy	community	2	0.61 (0.59, 0.63)	0.61 (0.59, 0.63)	8 (8, 8)
		31	21.74	52%	--	healthy	community	2	0.70 (0.70, 0.71)	0.70 (0.70, 0.71)	8 (8, 8)
Larson (unpublished-d)	25	122	70.8	53%	--	healthy	community	8	0.58 (0.50, 0.65)	0.58 (0.50, 0.65)	8 (8, 8)
McDonald et al. (in press)	26	89	34.15	36%	--	healthy	community	2	0.28 (0.11, 0.44)	0.54 (0.33, 0.76)	2 (2, 2)
Meyer et al. (2013)	27	43	19.14	44%	--	healthy	undergraduate	3	0.62 (0.48, 0.71)	0.57 (0.42, 0.66)	10 (10, 10)
Meyer et al. (2014)	28	43	12.74	34%	--	healthy	community	4	0.51 (0.39, 0.61)	0.44 (0.20, 0.68)	13 (6, 20)
Moser et al. (2019)	29	92	18.79	47%	--	healthy	undergraduate	2	0.77 (0.69, 0.86)	0.64 (0.53, 0.75)	16 (16, 16)
		102	21.02	57%	--	healthy	undergraduate	2	0.77 (0.70, 0.84)	0.58 (0.48, 0.68)	20 (20, 20)
		104	19.36	73%	--	healthy	undergraduate	2	0.73 (0.66, 0.79)	0.53 (0.45, 0.61)	19 (19, 19)
Moser (in prep)	30	162	35.33	62%	--	clinical high risk	community	6	0.74 (0.63, 0.80)	0.74 (0.63, 0.80)	8 (8, 8)
Muir et al. (2019)	31	128	20.62	53%	--	healthy	undergraduate	2	0.54 (0.49, 0.60)	0.61 (0.56, 0.67)	6 (6, 6)
Olson et al. (2018)	32	20	20.3	45%	--	healthy	undergraduate	1	0.85	0.88	6
		25	21	20%	mTBI	neurological	undergraduate	1	0.82	0.85	6
Olvet and Hajcak (2009)	33	53	--	62%	--	healthy	undergraduate	1	0.62	0.69	6
Pontifex et al. (2010)	34	56	9.6	43%	--	healthy	community	1	0.90	0.92	6
		57	19.9	60%	--	healthy	community	1	0.91	0.93	6
		26	65.7	46%	--	healthy	community	1	0.87	0.90	6
Rietdijk et al. (2014)	35	70	--	--	--	healthy	undergraduate	1	0.61	0.61	8

Sandre et al. (under review)	36	263	20.1	41%	--	healthy	undergraduate	12	0.68 (0.62, 0.77)	0.74 (0.69, 0.82)	6 (6, 6)
Seer et al. (2017)	37	13	63.15	23%	--	healthy	community	2	0.82 (0.80, 0.84)	0.82 (0.80, 0.84)	8 (8, 8)
		13	64.31	23%	Parkinson's Disease	neurological	community	2	0.68 (0.66, 0.71)	0.68 (0.66, 0.71)	8 (8, 8)
Singh et al. (2018)	38	28	69.2	39%	NA	healthy	community	2	0.38 (0.31, 0.45)	0.55 (0.47, 0.62)	4 (4, 4)
		28	69.8	39%	Parkinson's Disease	neurological	community	2	0.49 (0.39, 0.59)	0.65 (0.56, 0.74)	4 (4, 4)
Steele et al. (2016)	39	100	26.78	52%	--	healthy	community	1	0.81	0.94	2
Suchan et al. (2018)	40	14	23.5	64%	--	healthy	community	2	0.69 (0.69, 0.69)	0.53 (0.53, 0.53)	16 (16, 16)
Valadez and Simons (2018)	41	41	20.07	76%	--	healthy	undergraduate	1	0.62	0.72	5
Warren et al. (2020)	42	19	32.68	37%	--	healthy	community	2	0.82 (0.78, 0.87)	0.82 (0.78, 0.87)	8 (8, 8)
		16	30.69	38%	Gilles-de-la-Tourette Syndrome	neurological	community	2	0.76 (0.74, 0.77)	0.76 (0.74, 0.77)	8 (8, 8)
Xu and Inzlicht (2015)	43	12	18.09	33%	--	healthy	undergraduate	1	0.59	0.49	12

Note. The study number column provides an ID for each study. This ID is used to indicate which studies were excluded during the sensitivity analyses presented in Table 3. The ‘*n*’ column refers to the sample size of a given group. Some demographic information was missing for certain samples, and these samples were excluded from moderator analyses for the relevant missing characteristic. ‘# of Estimates’ refers to the number of internal consistency estimates that were obtained for a given sample. The ‘Original Reliability’ column shows the original coefficient alpha estimates for a given study. The ‘SB Reliability’ column shows the transformed coefficient alpha estimates for a given study using eight trials. Alpha estimates were transformed using the Spearman-Brown prophecy formula. The last three columns show the point estimates (mean) and range (minimum to maximum), when multiple internal consistency estimates were used from a given sample. Articles included in the meta-analysis are marked with an asterisk in the Reference section. All information for each sample and study is posted in the supplementary material on Open Science Framework. MDD = major depressive disorder, SUDs/AUDs = a mixed sample of participants with substance or alcohol use disorders, OCD = obsessive-compulsive disorder, GAD = generalized anxiety disorder, mTBI = mild traumatic brain injury

Table 2

Moderator Analyses for Coefficient Alpha Estimates of the Error-Related Negativity

Moderator	k_{study}	$k_{samples}$	$k_{estimates}$	n	$\hat{\alpha}_B$ (95% CI)	$\hat{\alpha}$ (95% CI)	t	p	Q_E (df)	p
Publication Status									1,442 (187)	< .001
Published	31	51	105	3,198	1.17 (0.99, 1.35)	.69 (.63, .74)	12.95	< .001	-	-
Unpublished	12	17	84	1,301	-.11 (-.44, 0.22)	.65 (.52, .75)	-.65	.52	-	-
Focal Outcome									1,139 (187)	< .001
Not Focal	29	47	139	2,299	1.07 (0.89, 1.25)	.66 (.59, .71)	11.76	< .001	-	-
Focal	14	21	50	2,200	0.20 (-.11, 0.52)	.72 (.62, .80)	1.28	.20	-	-
Paradigm[1]									1,484 (184)	< .001
Flanker	31	50	127	3,709	1.12 (0.95, 1.28)	.67 (.61, .72)	13.53	< .001	-	-
Go/NoGo	11	13	29	709	0.22 (0.09, 0.35)	.74 (.70, .77)	3.22	.002	-	-
Picture/Word Task	1	2	2	46	-.80 (-1.33, -.28)	.27 (.00, .57)	-3.01	.003	-	-
Simon	3	5	14	195	0.18 (-.40, 0.76)	.73 (.51, .85)	0.60	.55	-	-
Stroop	4	6	17	296	-.30 (-.70, 0.11)	.56 (.34, .71)	-1.44	.15	-	-
Clinical Status[1]									1,436 (185)	< .001
Healthy	42	51	143	3,703	1.16 (1.01, 1.31)	.69 (.63, .73)	15.15	< .001	-	-
Clinical High Risk	1	1	6	162	0.21 (-.73, 1.15)	.74 (.35, .90)	0.44	.66	-	-
Neurological	6	6	19	174	-.01 (-.28, 0.26)	.68 (.58, .76)	-.09	.93	-	-
Psychopathology	7	10	21	360	-.24 (-.47, -.01)	.60 (.50, .68)	-2.04	.04	-	-
EEG System[1]									967 (184)	< .001
BioSemi	13	17	33	1,058	1.10 (0.85, 1.35)	.67 (.57, .74)	8.64	< .001	-	-
ANT	4	5	8	222	0.12 (-.29, 0.54)	.71 (.56, .81)	0.58	.56	-	-
Brain Products	7	11	38	1,271	0.22 (-.23, 0.67)	.73 (.58, .83)	0.95	.34	-	-
EGI	15	25	88	1,426	-.07 (-.38, 0.25)	.65 (.52, .74)	-.42	.68	-	-
Neuroscan	6	10	22	422	0.11 (-.36, 0.58)	.70 (.52, .81)	0.45	.65	-	-
Age									1,530 (185)	< .001
Intercept	-	-	-	-	1.15 (0.93, 1.37)	.68 (.61, .75)	10.34	< .001	-	-
Age	41	66	187	4,376	-.00 (-.01, .01)	.68 (.68, .69)	-.07	.95	-	-

Percent Women									1,538 (186)	< .001
Intercept	-	-	-	-	1.30 (0.96, 1.64)	.73 (.62, .81)	7.58	< .001	-	-
Percent Women	42	67	188	4,429	-.32 (-.91, 0.28)	.63 (.32, .80)	-1.04	.30	-	-
Sample Type									1,524 (187)	< .001
Undergraduate	21	29	80	1,623	1.19 (0.98, 1.40)	.70 (.62, .75)	11.08	< .001	-	-
Community	23	39	109	2,776	-.09 (-.37, 0.18)	.67 (.56, .75)	-.66	.51	-	-
EEG Reference									1,437 (185)	< .001
Average Reference	23	38	126	2,666	1.10 (0.93, 1.27)	.67 (.60, .72)	12.42	< .001	-	-
Average Ear Lobes	3	4	7	174	-.09 (-.65, 0.48)	.64 (.36, .79)	-.31	.76	-	-
Average Mastoids	16	25	55	1,459	0.01 (-.27, 0.28)	.67 (.56, .75)	0.05	.96	-	-
Nose	1	1	1	100	1.79 (0.91, 2.68)	.95 (.87, .98)	3.99	< .001	-	-
Scoring Procedure[1]									1,494 (185)	< .001
Mean	37	57	110	4,066	1.11 (0.96, 1.26)	.67 (.62, .72)	14.56	< .001	-	-
Adaptive Mean	14	22	48	1,910	0.08 (0.01, 0.14)	.70 (.68, .72)	2.41	.02	-	-
Peak	9	13	25	945	0.13 (0.04, 0.21)	.71 (.68, .73)	2.93	.004	-	-
Peak-to-Peak	2	2	6	1,041	-.14 (-.22, -.06)	.62 (.59, .65)	-3.54	< .001	-	-
Length of Mean[1]									1,369 (156)	< .001
Intercept	-	-	-	-	1.14 (0.97, 1.32)	.68 (.62, .73)	13.15	< .001	-	-
Size of Mean	41	64	160	5,976	-.001 (-.002, 0.00)	.68 (.68, .68)	-1.16	.25	-	-
Sensors[1]									1,361 (187)	< .001
Single Sensor	33	50	108	3,695	1.17 (1.02, 1.32)	.69 (.64, .73)	15.38	< .001	-	-
Cluster of Sensors	21	36	81	2,006	-.08 (-.15, -.01)	.66 (.64, .69)	-2.31	.02	-	-
Trial Selection[1]									1,426 (187)	< .001
Initial	33	53	123	2,946	1.25 (1.08, 1.41)	.71 (.66, .76)	14.92	< .001	-	-
Random	28	41	66	2,727	-.25 (-.33, -.17)	.63 (.60, .66)	-6.00	< .001	-	-
Year of Publication									1,038 (106)	
Intercept	-	-	-	-	1.16 (0.65, 1.67)	.69 (.48, .81)	4.53	< .001	-	-
Year of Publication	30	49	108	3,145	0.003 (-.07, 0.07)	.69 (.67, .71)	0.10	.92	-	-
Ocular Artifact Correction									1,553 (187)	< .001

ICA	31	52	146	3,523	1.24 (1.07, 1.40)	.71 (.66, .75)	14.73	< .001	-	-
Regression	12	16	43	876	-.39 (-.71, -.06)	.57 (.41, .69)	-2.34	.02	-	-
Artifact Rejection									1,476 (185)	< .001
Automatic	30	49	135	3,410	1.18 (1.01, 1.36)	.69 (.64, .74)	13.50	< .001	-	-
None	1	2	8	46	-.97 (-1.90, -.04)	.19 (.00, .68)	-2.05	.04	-	-
Semiautomatic	10	14	36	804	-.15 (-.50, 0.20)	.64 (.50, .75)	-.85	.39	-	-
Visual	2	3	10	139	0.28 (-.40, 0.95)	.77 (.54, .88)	0.81	.42	-	-

Note. The first listed moderator of each set was entered as the intercept in the model. The Bonett-transformed coefficient alpha estimates ($\hat{\alpha}_B$) and their 95% confidence intervals (CIs) are shown for the intercept in the mixed model with each additional level showing the deviation from that intercept. The predicted coefficient alpha estimates ($\hat{\alpha}$) represent the back-transformed estimates. For ease of interpretation, each $\hat{\alpha}$ represents the estimate for that level of the moderator, rather than the deviation from the intercept. The sample size (n), number of alpha estimates ($k_{estimates}$), number of participant samples ($k_{samples}$), and number of studies (k_{study}) are shown for each level of a moderator. The Cochran's Q_E test for residual heterogeneity was used to determine whether the variability not accounted for by the moderator was larger than would be expected given the sampling variability alone. [1]Indicates moderator analysis wherein some studies or samples have estimates for more than one moderator. In such instances, the number of studies and/or samples for each moderator might be higher than the total studies and/or samples included in the meta-analysis due to overlap among levels. ICA = independent components analysis

Table 3

Sensitivity Analyses for Each Moderator of Coefficient Alpha Estimates of the Error-Related Negativity

Moderator	Excluded Estimates	k_{study}	$k_{samples}$	$k_{estimates}$	*n*	$\hat{\alpha}_B$ (95% CI)	$\hat{\alpha}$ (95% CI)	*t*	*p*	Q_E (*df*)	*p*
Publication Status	26, 39									1,308 (184)	< .001
Published		29	49	102	3,009	1.12 (0.96, 1.28)	.67 (.62, .72)	13.81	< .001	-	-
Unpublished		12	17	84	1,301	-.06 (-.35, 0.23)	.65 (.54, .74)	-.41	.68	-	-
Focal Outcome	15, 26, 35, 39									1,024 (182)	< .001
Not Focal		28	46	137	2,210	1.08 (0.91, 1.24)	.66 (.60, .71)	13.05	< .001	-	-
Focal		12	19	47	1,979	0.08 (-.22, 0.37)	.68 (.57, .77)	0.50	.62	-	-
Paradigm[1]	1, 5, 8, 9, 15, 26, 27, 28, 30, 39									1,132 (160)	< .001
Flanker		28	46	116	3,449	1.11 (0.95, 1.27)	.67 (.62, .72)	13.99	< .001	-	-
Go/NoGo		7	9	20	324	0.03 (-.26, 0.32)	.68 (.57, .76)	0.21	.84	-	-
Picture/Word Task		1	2	2	46	-.52 (-1.45, 0.41)	.44 (.00, .78)	-1.11	.27	-	-
Simon		3	5	11	195	0.08 (-.43, 0.59)	.70 (.49, .82)	0.32	.75		
Stroop		3	5	16	253	-.21 (-.71, 0.29)	.59 (.33, .75)	-.83	.41	-	-
Clinical Status[1]	2, 4, 12, 15, 18, 19, 21, 26, 30, 32, 38, 39, 42									1,204 (165)	< .001
Healthy		37	46	133	3,447	1.11 (0.99, 1.24)	.67 (.63, .71)	17.10	< .001	-	-
Clinical High Risk		1	1	5	162	0.33 (-.42, 1.07)	.76 (.50, .89)	0.86	.39	-	-
Neurological		4	4	14	121	0.02 (-.28, 0.32)	.68 (.57, .76)	0.14	.89	-	-
Psychopathology		6	8	17	253	-.37 (-.64, -.11)	.52 (.38, .64)	-2.77	.006	-	-
EEG System[1]	8, 26, 29, 34, 39, 43									689 (171)	< .001
BioSemi		11	15	30	854	1.03 (0.78, 1.27)	.64 (.54, .72)	8.20	< .001	-	-
ANT		2	3	5	118	0.19 (-.45, 0.83)	.70 (.44, .84)	0.59	.56	-	-

Brain Products		7	11	38	1,271	0.30 (-.09, 0.70)	.74 (.61, .82)	1.51	.13	-	-
EGI		14	24	86	1,324	-.02 (-.35, 0.30)	.63 (.49, .74)	-.14	.89	-	-
Neuroscan		5	8	17	276	0.17 (-.27, 0.61)	.70 (.53, .81)	0.75	.45	-	-
Age	15, 26, 34, 39									1,285 (177)	< .001
Intercept		-	-	-	-	1.04 (0.87, 1.21)	.65 (.58, .70)	12.29	< .001	-	-
Age		38	61	179	3,997	0.001 (-.007, 0.01)	.65 (.64, .65)	0.24	.81	-	-
Percent Women	15, 26, 29, 34, 39									1,303 (176)	< .001
Intercept		-	-	-	-	1.29 (0.96, 1.61)	.72 (.62, .80)	7.83	< .001	-	-
Percent Women		40	62	178	4,015	-.37 (-.95, 0.21)	.60 (.29, .78)	-1.25	.21	-	-
Sample Type	8, 12, 26, 39									1,351 (182)	< .001
Undergraduate		21	29	78	1,623	1.12 (0.94, 1.31)	.68 (.61, .73)	11.98	< .001	-	-
Community		21	37	106	2,587	-.06 (-.31, 0.19)	.65 (.56, .73)	-.47	.64	-	-
EEG Reference	8, 13, 15, 26, 43									1,374 (179)	< .001
Average Reference		23	38	125	2,666	1.09 (0.91, 1.26)	.66 (.60, .72)	12.24	< .001	-	-
Average Ear Lobes		2	3	5	162	0.07 (-.55, 0.69)	.69 (.41, .83)	0.22	.82	-	-
Average Mastoids		15	24	52	1,319	0.02 (-.27, 0.30)	.67 (.56, .75)	0.10	.92	-	-
Nose[2]		0	0	0	-	-	-	-	-	-	-
Scoring Procedure[1]	1, 8, 14, 20, 21, 23, 25, 26, 29, 30, 36, 39									875 (140)	< .001
Mean		33	51	89	3,417	1.06 (0.92, 1.20)	.65 (.60, .70)	15.04	< .001	-	-
Adaptive Mean		13	21	41	1,748	0.05 (-.03, 0.14)	.67 (.64, .70)	1.27	.21	-	-
Peak		6	8	13	282	0.22 (-.06, 0.50)	.72 (.63, .79)	1.53	.13	-	-
Peak-to-Peak		1	1	1	263	-.15 (-.41, 0.11)	.60 (.48, .69)	-1.13	.26	-	-
Length of Mean[1]	1, 12, 14, 20, 21, 23, 25, 30, 39									849 (134)	< .001
Intercept		-	-	-	-	1.05 (0.88, 1.21)	.65 (.59, .70)	12.60	< .001	-	-

Size of Mean		39	62	136	5,525	0.00 (-.001, 0.001)	.65 (.65, .65)	.02	.99	-	-
Sensors[1]	2, 9, 20, 21, 23, 25, 26, 36, 39									1,032 (161)	< .001
Single Sensor		30	46	93	3,177	1.10 (0.96, 1.24)	.67 (.62, .71)	15.65	< .001	-	-
Cluster of Sensors		21	34	70	1,700	0.01 (-.09, 0.11)	.67 (.63, .70)	0.13	.90	-	-
Trial Selection[1]	1, 8, 9, 10, 20, 21, 25, 26, 39									1,011 (164)	< .001
Initial		32	51	114	2,797	1.19 (1.04, 1.33)	.70 (.65, .74)	16.25	< .001	-	-
Random		24	37	52	2,354	-.22 (-.32, -.11)	.62 (.58, .66)	-4.03	< .001	-	-
Year of Publication	4									937 (105)	< .001
Intercept		-	-	-	-	1.12 (0.69, 1.55)	.67 (.50, .79)	5.17	< .001	-	-
Year of Publication		29	48	107	3,045	0.001 (-.06, 0.06)	.67 (.65, .69)	0.02	.98	-	-
Ocular Artifact Correction	15, 26, 33, 39									1,413 (182)	< .001
ICA		30	51	145	3,423	1.18 (1.04, 1.33)	.69 (.65, .74)	16.08	< .001	-	-
Regression		10	14	39	683	-.39 (-.70, -.09)	.55 (.38, .67)	-2.56	.01	-	-
Artifact Rejection	8, 9, 15, 26, 39									1,287 (177)	< .001
Automatic		28	47	132	3,221	1.13 (0.98, 1.28)	.68 (.63, .72)	14.96	< .001	-	-
None		1	2	8	46	-.92 (-1.69, -.14)	.19 (.00, .63)	-2.34	.02	-	-
Semiautomatic		10	14	35	753	-.13 (-.42, 0.17)	.63 (.51, .73)	-.84	.40	-	-
Visual		2	3	10	139	0.23 (-.35, 0.81)	.74 (.54, .86)	0.79	.43	-	-

Note. Sensitivity analyses mirror the results presented in Table 2, with the exception that influential estimates were removed from these moderator analyses. The ID for the studies with estimates removed from analyses are shown in the 'Excluded Estimates' column, and th study corresponding to that ID can be found in Table 1. When a study had multiple estimates, it is possible that some estimates, but not others, were included in the sensitivity analyses. The first listed moderator of each set was entered as the intercept in the model. The Bonett-transformed coefficient alpha estimates ($\widehat{\alpha}_B$) and their 95% confidence intervals (CIs) are shown for the intercept in the mixed model with each additional level showing the deviation from that intercept. The predicted coefficient alpha estimates ($\widehat{\alpha}$) represent the back-transformed estimates. For ease of interpretation, each $\hat{\alpha}$ represents the estimate for that level of the moderator, rather than the deviation from the intercept. The sample size (*n*), number of alpha estimates ($k_{estimates}$), number of participant samples ($k_{samples}$), and number of studies (k_{study}) are shown for each level of a moderator. The Cochran's Q_E test for residual heterogeneity was used to determine whether the variability not accounted for by the moderator was larger than would be expected given the sampling variability alone. [1]Indicates moderator analysis wherein some samples have estimates for more than one moderator. In such instances, the number of studies and/or samples for each moderator might be higher than the total studies and/or samples included in the meta-analysis due to overlap among levels. [2] There were no levels left in the dataset after excluding influential estimates. ICA = independent components analysis

Figure Captions

Figure 1. PRISMA Flow Diagram

Figure 2. Forest plot of the coefficient alpha point estimates and 95% confidence intervals for the first half of all estimates included in the reliability generalization study. The estimate for the random effects (RE) intercept-only model for all included studies from Figures 2 and 3 is shown at the bottom. A dotted line is shown for the lower limit of coefficient alpha, the value of the summary estimate (.68), and the upper limit of coefficient alpha, respectively.

Figure 3. Forest plot of the coefficient alpha point estimates and 95% confidence intervals for the second half of all estimates included in the reliability generalization study. The estimate for the random effects (RE) intercept-only model for all included studies from Figures 2 and 3 is shown at the bottom. A dotted line is shown for the lower limit of coefficient alpha, the value of the summary estimate (.68), and the upper limit of coefficient alpha, respectively.

Figure 4. Line plots showing the relationship between the numbers of trials (four to twenty trials) used for computing internal consistency estimates and the estimated internal consistency ($\widehat{\alpha}$) using an intercept-only random effects model. The plot on the top (A) uses all estimates from the meta-analysis. The plot on the bottom (B) excludes influential estimates from the model. Shaded areas represent 95% confidence intervals. The dotted line shows the internal consistency at .70, and the dashed line shows internal consistency at .80.

Figure 1

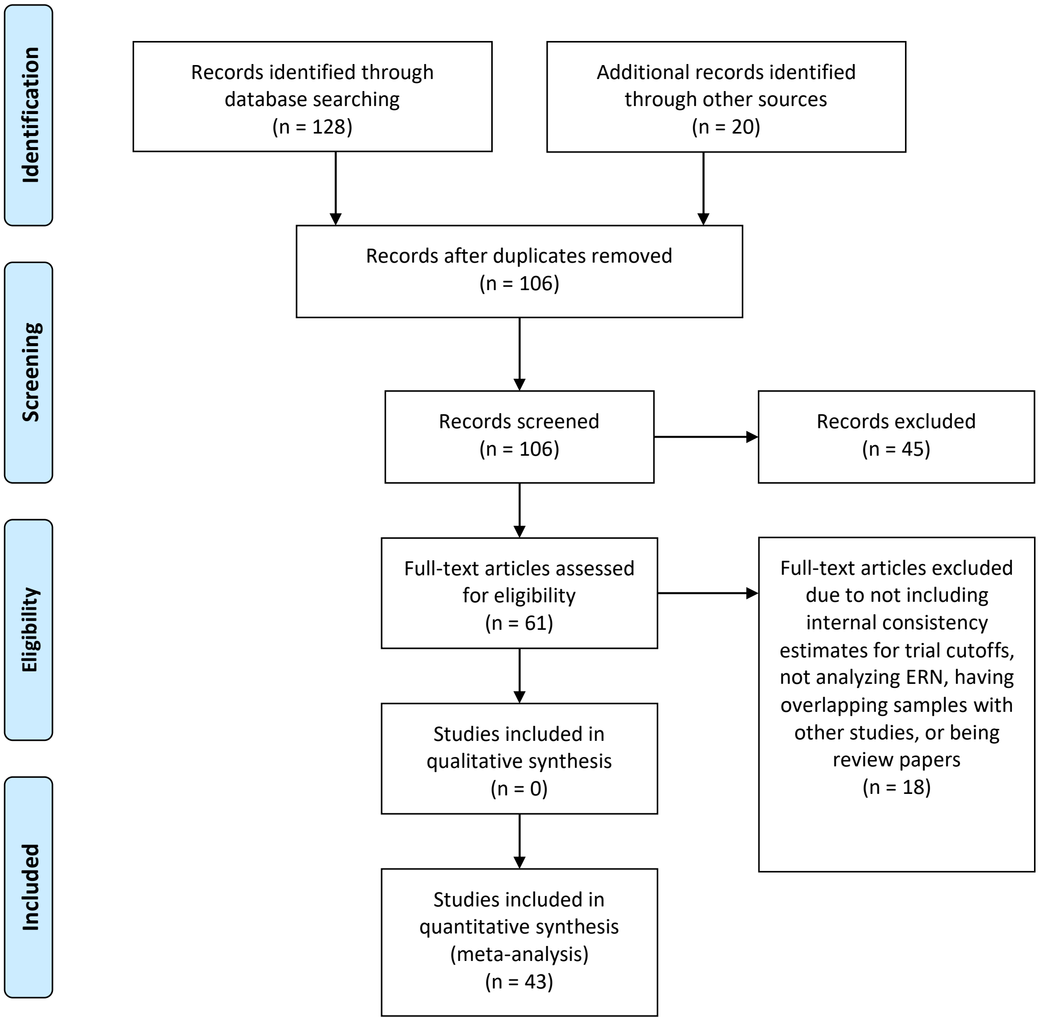

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 128)
Additional records identified through other sources (n = 20)
Records after duplicates removed (n = 106)
Records screened (n = 106)
Records excluded (n = 45)
Full-text articles assessed for eligibility (n = 61)
Full-text articles excluded due to not including internal consistency estimates for trial cutoffs, not analyzing ERN, having overlapping samples with other studies, or being review papers (n = 18)
Studies included in qualitative synthesis (n = 0)
Studies included in quantitative synthesis (meta-analysis) (n = 43)

Figure 2

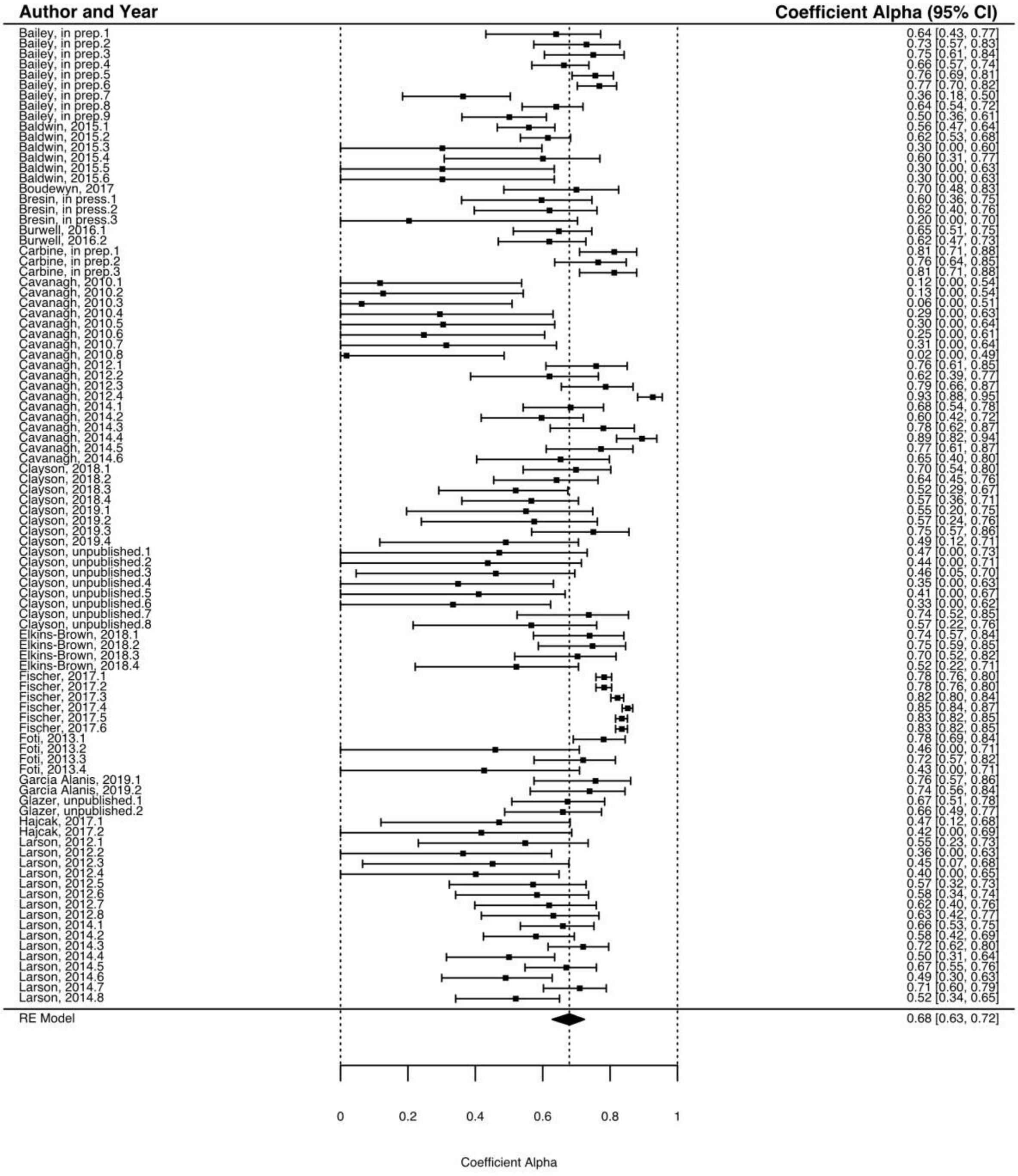

Author and Year
Coefficient Alpha (95% CI)
Bailey, in prep.1 0.64 [0.43, 0.77]
Bailey, in prep.2 0.73 [0.57, 0.83]
Bailey, in prep.3 0.75 [0.61, 0.84]
Bailey, in prep.4 0.66 [0.57, 0.74]
Bailey, in prep.5 0.76 [0.69, 0.81]
Bailey, in prep.6 0.77 [0.70, 0.82]
Bailey, in prep.7 0.36 [0.18, 0.50]
Bailey, in prep.8 0.64 [0.54, 0.72]
Bailey, in prep.9 0.50 [0.36, 0.61]
Baldwin, 2015.1 0.56 [0.47, 0.64]
Baldwin, 2015.2 0.62 [0.53, 0.68]
Baldwin, 2015.3 0.30 [0.00, 0.60]
Baldwin, 2015.4 0.60 [0.31, 0.77]
Baldwin, 2015.5 0.30 [0.00, 0.63]
Baldwin, 2015.6 0.30 [0.00, 0.63]
Boudewyn, 2017 0.70 [0.48, 0.83]
Bresin, in press.1 0.60 [0.36, 0.75]
Bresin, in press.2 0.62 [0.40, 0.76]
Bresin, in press.3 0.20 [0.00, 0.70]
Burwell, 2016.1 0.65 [0.51, 0.75]
Burwell, 2016.2 0.62 [0.47, 0.73]
Carbine, in prep.1 0.81 [0.71, 0.88]
Carbine, in prep.2 0.76 [0.64, 0.85]
Carbine, in prep.3 0.81 [0.71, 0.88]
Cavanagh, 2010.1 0.12 [0.00, 0.54]
Cavanagh, 2010.2 0.13 [0.00, 0.54]
Cavanagh, 2010.3 0.06 [0.00, 0.51]
Cavanagh, 2010.4 0.29 [0.00, 0.63]
Cavanagh, 2010.5 0.30 [0.00, 0.64]
Cavanagh, 2010.6 0.25 [0.00, 0.61]
Cavanagh, 2010.7 0.31 [0.00, 0.64]
Cavanagh, 2010.8 0.02 [0.00, 0.49]
Cavanagh, 2012.1 0.76 [0.61, 0.85]
Cavanagh, 2012.2 0.62 [0.39, 0.77]
Cavanagh, 2012.3 0.79 [0.66, 0.87]
Cavanagh, 2012.4 0.93 [0.88, 0.95]
Cavanagh, 2014.1 0.68 [0.54, 0.78]
Cavanagh, 2014.2 0.60 [0.42, 0.72]
Cavanagh, 2014.3 0.78 [0.62, 0.87]
Cavanagh, 2014.4 0.89 [0.82, 0.94]
Cavanagh, 2014.5 0.77 [0.61, 0.87]
Cavanagh, 2014.6 0.65 [0.40, 0.80]
Clayson, 2018.1 0.70 [0.54, 0.80]
Clayson, 2018.2 0.64 [0.45, 0.76]
Clayson, 2018.3 0.52 [0.29, 0.67]
Clayson, 2018.4 0.57 [0.36, 0.71]
Clayson, 2019.1 0.55 [0.20, 0.75]
Clayson, 2019.2 0.57 [0.24, 0.76]
Clayson, 2019.3 0.75 [0.57, 0.86]
Clayson, 2019.4 0.49 [0.12, 0.71]
Clayson, unpublished.1 0.47 [0.00, 0.73]
Clayson, unpublished.2 0.44 [0.00, 0.71]
Clayson, unpublished.3 0.46 [0.05, 0.70]
Clayson, unpublished.4 0.35 [0.00, 0.63]
Clayson, unpublished.5 0.41 [0.00, 0.67]
Clayson, unpublished.6 0.33 [0.00, 0.62]
Clayson, unpublished.7 0.74 [0.52, 0.85]
Clayson, unpublished.8 0.57 [0.22, 0.76]
Elkins-Brown, 2018.1 0.74 [0.57, 0.84]
Elkins-Brown, 2018.2 0.75 [0.59, 0.85]
Elkins-Brown, 2018.3 0.70 [0.52, 0.82]
Elkins-Brown, 2018.4 0.52 [0.22, 0.71]
Fischer, 2017.1 0.78 [0.76, 0.80]
Fischer, 2017.2 0.78 [0.76, 0.80]
Fischer, 2017.3 0.82 [0.80, 0.84]
Fischer, 2017.4 0.85 [0.84, 0.87]
Fischer, 2017.5 0.83 [0.82, 0.85]
Fischer, 2017.6 0.83 [0.82, 0.85]
Foti, 2013.1 0.78 [0.69, 0.84]
Foti, 2013.2 0.46 [0.00, 0.71]
Foti, 2013.3 0.72 [0.57, 0.82]
Foti, 2013.4 0.43 [0.00, 0.71]
García Alanis, 2019.1 0.76 [0.57, 0.86]
García Alanis, 2019.2 0.74 [0.56, 0.84]
Glazer, unpublished.1 0.67 [0.51, 0.78]
Glazer, unpublished.2 0.66 [0.49, 0.77]
Hajcak, 2017.1 0.47 [0.12, 0.68]
Hajcak, 2017.2 0.42 [0.00, 0.69]
Larson, 2012.1 0.55 [0.23, 0.73]
Larson, 2012.2 0.36 [0.00, 0.63]
Larson, 2012.3 0.45 [0.07, 0.68]
Larson, 2012.4 0.40 [0.00, 0.65]
Larson, 2012.5 0.57 [0.32, 0.73]
Larson, 2012.6 0.58 [0.34, 0.74]
Larson, 2012.7 0.62 [0.40, 0.76]
Larson, 2012.8 0.63 [0.42, 0.77]
Larson, 2014.1 0.66 [0.53, 0.75]
Larson, 2014.2 0.58 [0.42, 0.69]
Larson, 2014.3 0.72 [0.62, 0.80]
Larson, 2014.4 0.50 [0.31, 0.64]
Larson, 2014.5 0.67 [0.55, 0.76]
Larson, 2014.6 0.49 [0.30, 0.63]
Larson, 2014.7 0.71 [0.60, 0.79]
Larson, 2014.8 0.52 [0.34, 0.65]
RE Model 0.68 [0.63, 0.72]
0
0.2
0.4
0.6
0.8
1
Coefficient Alpha

Figure 3

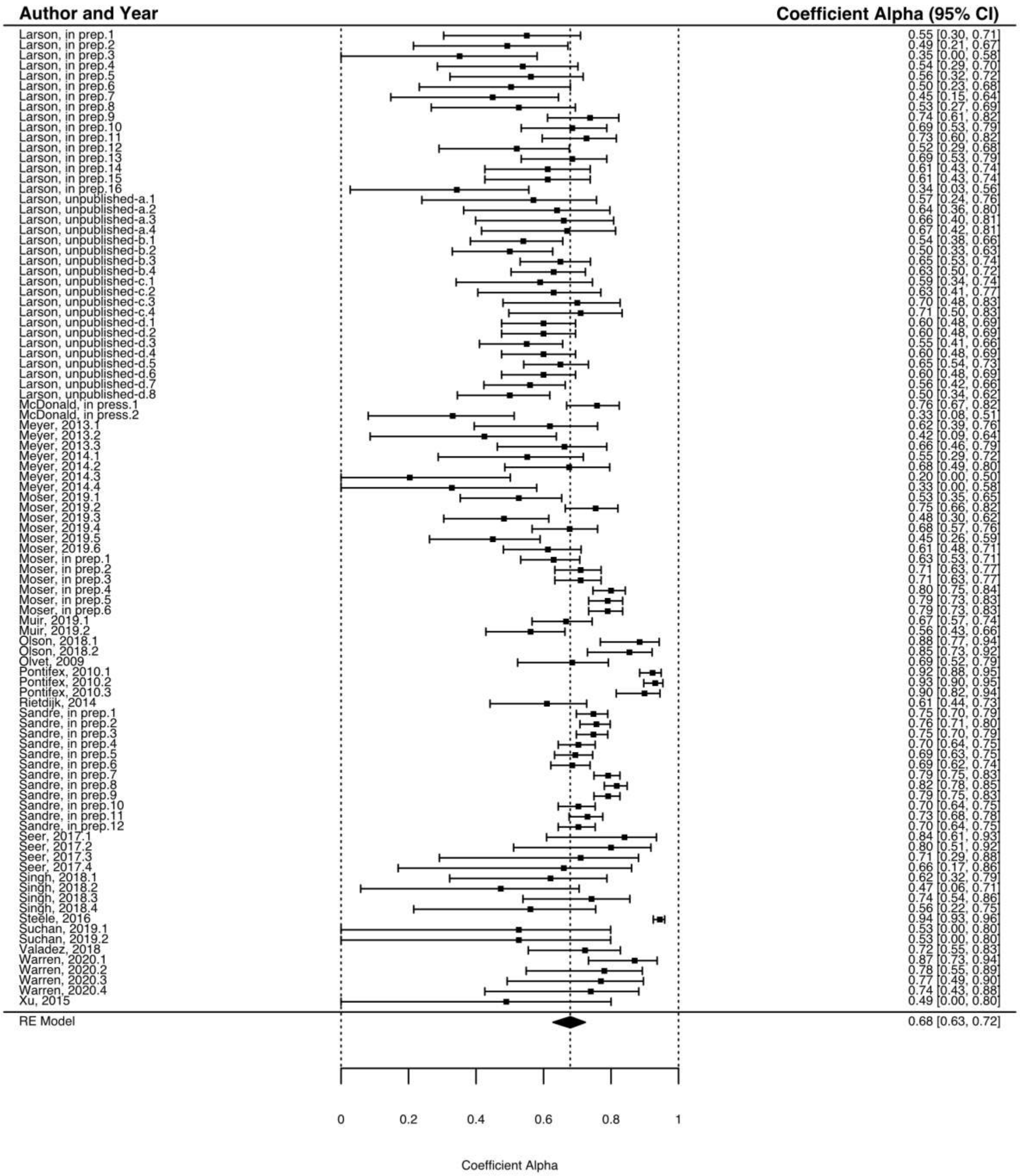
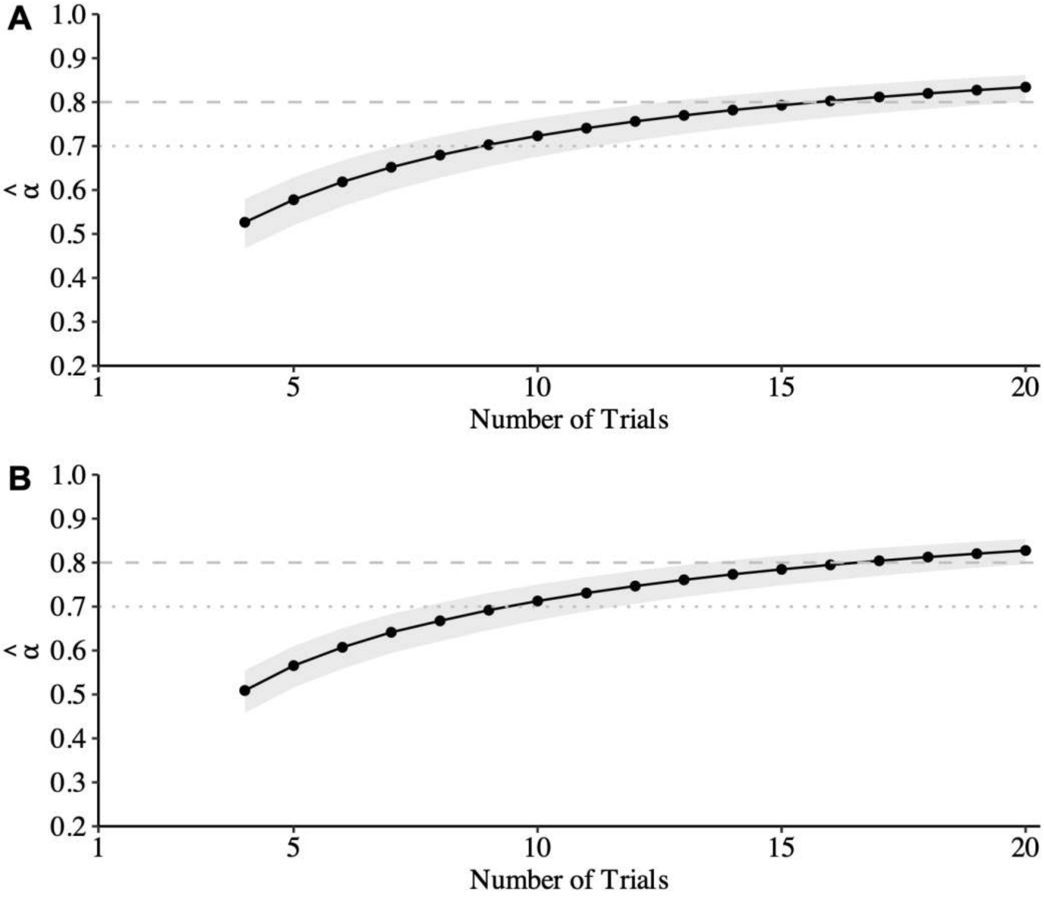

Author and Year
Coefficient Alpha (95% CI)
Larson, in prep.1 0.55 [0.30, 0.71]
Larson, in prep.2 0.49 [0.21, 0.67]
Larson, in prep.3 0.35 [0.00, 0.58]
Larson, in prep.4 0.54 [0.29, 0.70]
Larson, in prep.5 0.56 [0.32, 0.72]
Larson, in prep.6 0.50 [0.23, 0.68]
Larson, in prep.7 0.45 [0.15, 0.64]
Larson, in prep.8 0.53 [0.27, 0.69]
Larson, in prep.9 0.74 [0.61, 0.82]
Larson, in prep.10 0.69 [0.53, 0.79]
Larson, in prep.11 0.73 [0.60, 0.82]
Larson, in prep.12 0.52 [0.29, 0.68]
Larson, in prep.13 0.69 [0.53, 0.79]
Larson, in prep.14 0.61 [0.43, 0.74]
Larson, in prep.15 0.61 [0.43, 0.74]
Larson, in prep.16 0.34 [0.03, 0.56]
Larson, unpublished-a.1 0.57 [0.24, 0.76]
Larson, unpublished-a.2 0.64 [0.36, 0.80]
Larson, unpublished-a.3 0.66 [0.40, 0.81]
Larson, unpublished-a.4 0.67 [0.42, 0.81]
Larson, unpublished-b.1 0.54 [0.38, 0.66]
Larson, unpublished-b.2 0.50 [0.33, 0.63]
Larson, unpublished-b.3 0.65 [0.53, 0.74]
Larson, unpublished-b.4 0.63 [0.50, 0.72]
Larson, unpublished-c.1 0.59 [0.34, 0.74]
Larson, unpublished-c.2 0.63 [0.41, 0.77]
Larson, unpublished-c.3 0.70 [0.48, 0.83]
Larson, unpublished-c.4 0.71 [0.50, 0.83]
Larson, unpublished-d.1 0.60 [0.48, 0.69]
Larson, unpublished-d.2 0.60 [0.48, 0.69]
Larson, unpublished-d.3 0.55 [0.41, 0.66]
Larson, unpublished-d.4 0.60 [0.48, 0.69]
Larson, unpublished-d.5 0.65 [0.54, 0.73]
Larson, unpublished-d.6 0.60 [0.48, 0.69]
Larson, unpublished-d.7 0.56 [0.42, 0.66]
Larson, unpublished-d.8 0.50 [0.34, 0.62]
McDonald, in press.1 0.76 [0.67, 0.82]
McDonald, in press.2 0.33 [0.08, 0.51]
Meyer, 2013.1 0.62 [0.39, 0.76]
Meyer, 2013.2 0.42 [0.09, 0.64]
Meyer, 2013.3 0.66 [0.46, 0.79]
Meyer, 2014.1 0.55 [0.29, 0.72]
Meyer, 2014.2 0.68 [0.49, 0.80]
Meyer, 2014.3 0.20 [0.00, 0.50]
Meyer, 2014.4 0.33 [0.00, 0.58]
Moser, 2019.1 0.53 [0.35, 0.65]
Moser, 2019.2 0.75 [0.66, 0.82]
Moser, 2019.3 0.48 [0.30, 0.62]
Moser, 2019.4 0.68 [0.57, 0.76]
Moser, 2019.5 0.45 [0.26, 0.59]
Moser, 2019.6 0.61 [0.48, 0.71]
Moser, in prep.1 0.63 [0.53, 0.71]
Moser, in prep.2 0.71 [0.63, 0.77]
Moser, in prep.3 0.71 [0.63, 0.77]
Moser, in prep.4 0.80 [0.75, 0.84]
Moser, in prep.5 0.79 [0.73, 0.83]
Moser, in prep.6 0.79 [0.73, 0.83]
Muir, 2019.1 0.67 [0.57, 0.74]
Muir, 2019.2 0.56 [0.43, 0.66]
Olson, 2018.1 0.88 [0.77, 0.94]
Olson, 2018.2 0.85 [0.73, 0.92]
Olvet, 2009 0.69 [0.52, 0.79]
Pontifex, 2010.1 0.92 [0.88, 0.95]
Pontifex, 2010.2 0.93 [0.90, 0.95]
Pontifex, 2010.3 0.90 [0.82, 0.94]
Rietdijk, 2014 0.61 [0.44, 0.73]
Sandre, in prep.1 0.75 [0.70, 0.79]
Sandre, in prep.2 0.76 [0.71, 0.80]
Sandre, in prep.3 0.75 [0.70, 0.79]
Sandre, in prep.4 0.70 [0.64, 0.75]
Sandre, in prep.5 0.69 [0.63, 0.75]
Sandre, in prep.6 0.69 [0.62, 0.74]
Sandre, in prep.7 0.79 [0.75, 0.83]
Sandre, in prep.8 0.82 [0.78, 0.85]
Sandre, in prep.9 0.79 [0.75, 0.83]
Sandre, in prep.10 0.70 [0.64, 0.75]
Sandre, in prep.11 0.73 [0.68, 0.78]
Sandre, in prep.12 0.70 [0.64, 0.75]
Seer, 2017.1 0.84 [0.61, 0.93]
Seer, 2017.2 0.80 [0.51, 0.92]
Seer, 2017.3 0.71 [0.29, 0.88]
Seer, 2017.4 0.66 [0.17, 0.86]
Singh, 2018.1 0.62 [0.32, 0.79]
Singh, 2018.2 0.47 [0.06, 0.71]
Singh, 2018.3 0.74 [0.54, 0.86]
Singh, 2018.4 0.56 [0.22, 0.75]
Steele, 2016 0.94 [0.93, 0.96]
Suchan, 2019.1 0.53 [0.00, 0.80]
Suchan, 2019.2 0.53 [0.00, 0.80]
Valadez, 2018 0.72 [0.55, 0.83]
Warren, 2020.1 0.87 [0.73, 0.94]
Warren, 2020.2 0.78 [0.55, 0.89]
Warren, 2020.3 0.77 [0.49, 0.90]
Warren, 2020.4 0.74 [0.43, 0.88]
Xu, 2015 0.49 [0.00, 0.80]
RE Model 0.68 [0.63, 0.72]
0
0.2
0.4
0.6
0.8
1
Coefficient Alpha

Figure 4


A
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
$\hat{\alpha}$
1
5
10
15
20
Number of Trials
B
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
$\hat{\alpha}$
1
5
10
15
20
Number of Trials