Journal of Abnormal Psychology
2017, Vol. 126, No. 6, 823–834

© 2017 American Psychological Association
0021-843X/17/$12.00 http://dx.doi.org/10.1037/abn0000274

Psychometrics and the Neuroscience of Individual Differences: Internal Consistency Limits Between-Subjects Effects

Greg Hajcak
Stony Brook University

Alexandria Meyer
Florida State University

Roman Kotov
Stony Brook University

In the clinical neuroscience literature, between-subjects differences in neural activity are presumed to reflect reliable measures—even though the psychometric properties of neural measures are almost never reported. The current article focuses on the critical importance of assessing and reporting internal consistency reliability—the homogeneity of "items" that comprise a neural "score." We demonstrate how variability in the internal consistency of neural measures limits between-subjects (i.e., individual differences) effects. To this end, we utilize error-related brain activity (i.e., the error-related negativity or ERN) in both healthy and generalized anxiety disorder (GAD) participants to demonstrate options for psychometric analyses of neural measures; we examine between-groups differences in internal consistency, between-groups effect sizes, and between-groups discriminability (i.e., ROC analyses)—all as a function of increasing items (i.e., number of trials). Overall, internal consistency should be used to inform experimental design and the choice of neural measures in individual differences research. The internal consistency of neural measures is necessary for interpreting results and guiding progress in clinical neuroscience—and should be routinely reported in all individual differences studies.

General Scientific Summary
Limited advances from clinical neuroscience may stem from a failure in the field to consider basic measurement properties of neural measures. This article highlights the impact of internal consistency on between-groups effect sizes, and suggests reporting and using internal consistency to interpret and guide individual differences research that uses neuroscience measures.

Keywords: psychometrics, neuroscience, reliability, internal consistency

Neuroscience has made impressive strides in understanding how the human brain works—advances that largely come from *within-subjects* studies. Typically, brain activity is quantified during two or more experimental conditions in each participant, and these values are compared. Through sophisticated experimental design and subtraction techniques, these within-subjects differences shed light on the functional significance of neural activity implicated in specific psychological processes.

Throughout the "decade of the brain" and since, clinical psychologists and psychiatrists have attempted to utilize neuroscientific methods to better understand mental illness. A common strategy for this line of inquiry has been to compare brain activity between a group of individuals with a specific diagnosis (e.g., major depressive disorder) to a group of individuals with no history of any psychiatric disease (i.e., healthy control participants). In such studies, differences between groups have been interpreted in terms of the "pathophysiology" or "neural basis" of a specific psychological disorder or trait. This type of reductionism is rife with serious conceptual and philosophical difficulties (Miller, 2010); practically, this approach has produced relatively few significant advances in understanding, treating, and preventing mental illness (Insel et al., 2010; Stringaris, 2015).

The recent Research Domain Criteria (RDoC) initiative of the National Institute of Mental Health suggests that neuroscientific advances in mental illness have been hampered by the categorical system used to characterize psychopathology (Cuthbert & Insel, 2013; Cuthbert & Kozak, 2013; Insel et al., 2010; Kozak & Cuthbert, 2016; Sanislow et al., 2010). Specifically, the *Diagnostic and Statistical Manual* (*DSM*) and its variants conceptualize disorders in terms of a group of polythetic criteria; as a result, diagnoses are characterized by significant heterogeneity. For instance, two individuals can share the same diagnosis and have no symptoms in common. In addition, comorbidity is the rule rather than the exception (Kessler et al., 1994), which further increases heterogeneity within diagnostic groups. In

This article was published Online First April 27, 2017.

Greg Hajcak, Department of Psychology, Stony Brook University; Alexandria Meyer, Department of Psychology, Florida State University; Roman Kotov, Department of Psychiatry, Stony Brook University.

Correspondence concerning this article should be addressed to Greg Hajcak, Department of Psychology, Stony Brook University, Stony Brook, NY 11794. E-mail: greg.hajcak@stonybrook.edu

these ways, neuroscientific progress in psychopathology may have been hampered by nosology itself.

These concerns about *DSM*-based clinical diagnoses, though not exhaustive, have to do with *validity* and whether *DSM*-based diagnostic categories accurately reflect divisions between mental illness and health, and between different forms of mental illness. Rather than hoping to find neural correlates of constructs (i.e., diagnoses) with questionable validity, the RDoC project proposes an alternative approach: begin by focusing on neurobehavioral constructs with well-defined neural circuits that likely relate to *continuous* variability in functioning (e.g., symptoms) that cut across traditional diagnostic boundaries (Cuthbert & Kozak, 2013; Kozak & Cuthbert, 2016; Sanislow et al., 2010). RDoC is an effort to ground psychopathology research in neuroscience, and the RDoC matrix is analogous to a proposed periodic table of elements: rows represent processes that can be studied across various units of analysis (i.e., columns). The hope is that once the RDoC matrix is filled out, it will suggest a neuroscientifically informed way forward for conceptualizing psychopathology and improving our understanding and treatment of mental illness.

By emphasizing processes linked to fundamental neural systems, RDoC reflects an effort to build a science of psychopathology from a neuroscientific knowledge base rather than the purely descriptive approach of the *DSM–5*; this tactic is intended to bolster validity of the system from the perspective of pathophysiology. However, unlike the majority of studies that have shed light on how the brain works using *within-subjects* studies, understanding mental illness and individual differences using neuroscientific methods is an effort to understand *between-subjects* variability in within-subjects effects. The RDoC matrix is, implicitly, a matrix of processes and measures that are presumed to vary in the population and explain meaningful *individual differences* in clinically relevant behaviors (Hajcak & Patrick, 2015; Patrick & Hajcak, 2016). As an example, cognitive or affective neuroscience has shed light on neural systems implicated in the acquisition and expression of fear using within-subjects studies—to oversimplify, it seems quite clear that the amygdala is implicated in the acquisition and expression of fear-related behaviors (Tovote, Fadok, & Lüthi, 2015). However, whether variation in amygdala function relates to *individual differences* in the acquisition and expression of fear is a question about between-subjects variation in within-subjects effects.

Within-subjects comparisons only deal with means and *SD*s in one condition versus another aggregated across all participants—from the standpoint of psychometrics, a within-subject difference (i.e., a difference between two conditions) does not ensure high reliability, in the psychometric sense of the word. The clinical neuroscience literature is replete with examples where tasks are described as "reliably" eliciting certain patterns of neural activity; unfortunately, most of these studies actually mean "robustly." For instance, fear-eliciting stimuli *robustly* activate the amygdala. Across many studies, the average amygdala response to fearful faces is larger than the average response to neutral faces; that is, there is an observed mean-level difference across conditions that has been reported in many samples (Phan, Wager, Taylor, & Liberzon, 2002). The replicability and robustness of within-subjects (i.e., mean-level) differences does not, however, address the suitability of amygdala activation *as an individual difference measure*. It is possible that the amygdala responds differentially to fearful and neutral stimuli, and that the magnitude of this effect varies widely from trial to trial, or across testing sessions. Robust within-subjects effects do not necessarily imply good measures for studying individual differences. The potential lack of relationship between within- and between-subjects effects may reflect conceptual or mechanistic misunderstandings; however, divergence between within- and between-subjects relationships can reflect measurement issues. For a neural measure like amygdala activation to be appropriate as a measure of individual differences, it has to be *reliable.*

Reliability refers to the *consistency of scores on a measure*—in neuroscience, this would amount to the tendency for *individuals* to show similar scores or values across repeated measurements of neural activity. The similarity of scores can be evaluated within (i.e., split-half reliability) or across (i.e., test–retest reliability) testing sessions. Regardless, each person's score is assumed to reflect both their true score on a measure, and error. True score refers to the systematic variation common to test items that underlies individual differences among participants; error reflects any variation that cannot be reproduced—and would include variability associated with specific items, methods, or administrations. Studying between-subjects variability (i.e., individual differences) requires *reliable* measures. A measure cannot be valid if it is not reliable: reliability is a *prerequisite* for validity (Cronbach & Meehl, 1955; Meehl, 1995, 1986).[1]

From a psychometric standpoint, the lackluster progress of clinical neuroscience may have to do with two issues, in isolation or combination: diagnostic categories that have questionable validity, or alternatively, a wealth of studies have used neural measures with inadequate reliability. The former possibility has motivated an entire shift in National Institutes of Mental Health (NIMH) funding priorities (i.e., the RDoC enterprise); the latter possibility has not been considered sufficiently. One *possibility* is that many neural measures have inadequate reliability to function well as measures of individual differences. This possibility can only be addressed empirically: researchers must evaluate and present psychometric properties of neural measures that are being treated as individual difference variables, much the way the reliabilities of more traditional self-report measures are routinely presented in publications.

Concerns about the psychometric properties of neural measures are not specific to the RDoC initiative. Any neuroscientific study of psychopathology—even in relation to *DSM*-based disorders—requires valid and reliable neural measures. For instance, major classes of mental disorders appear to be substantially heritable, although specifying the pathway from genetic liability to the expression of that risk has proven exceedingly difficult. One potential strategy is to identify *endophenotypes* for psychiatric disease. Endophenotypes are unobservable *traits* that mediate the association between genetic risk and the expression of a given phenotype (Gottesman & Gould, 2003; Kendler & Neale, 2010). The potential importance of endophenotypes comes from the fact that endophenotypes are less complex than associated disease

[1] It is important to note that reliability is necessary, but not sufficient, for validity. That is, an individual difference measure can be reliable, but not valid. Although we focus here on reliability, we further consider its relationship with validity in the discussion.

states, and might, therefore, be more amenable to genetic analyses. However, as Lilienfeld (2014) points out, it is possible that "endophenotypic markers based on single laboratory tasks may possess substantial amounts of situational uniqueness and therefore high levels of measurement error" (p. 135). That is, poor psychometric properties would limit the potential utility of endophenotypes. Indeed, the association between any two individual difference variables will be underestimated if either has low reliability—this is true for self-report, behavioral, and biological measures of individual differences (Rodebaugh et al., 2016).

Psychometric Properties of Neural Measures

The psychometric properties of most neural measures that have been examined in relation to psychopathology and other individual difference variables have not been evaluated sufficiently (Hajcak & Patrick, 2015; Lilienfeld, 2014). Broadly, the clinical neuroscience literature has used multiple neuroimaging modalities to index individual differences in neural function, including but not limited to metrics derived from functional magnetic resonance imaging (fMRI), positron emission tomography (PET), electroencephalography (EEG), event-related brain potentials (ERPs), and magnetoencephalography (MEG). Each of these neuroimaging modalities quantifies different properties of neural activation (i.e., compensatory hemodynamic response after neural activity in the case of fMRI, electrical and magnetic fields generated by neural activity in the case of ERP and MEG, respectively). Neural measures derived from each of these methods, therefore, are associated with different sources and amounts of noise. For instance, blinks and ocular movements have a larger impact on EEG/MEG data, whereas physical movement has a larger impact on fMRI data.

In the current article we focus on the impact of internal consistency reliability on the ability of neural measures to relate to individual differences. As a demonstration, we use a specific neural measure: error-related brain activity measured using ERPs (i.e., the error-related negativity or ERN). We would note, however, that error-related brain activity has also been quantified using other neural measures—and that *none* of the psychometric issues discussed in the current article are specific to error-related brain activity or to neural activity measured using ERPs. Rather, we focus on ERN because there have been a relatively large number of studies linking ERN to individual differences in psychopathology (Cavanagh & Shackman, 2015; Moser, Moran, Schroder, Donnellan, & Yeung, 2013), as well as several articles from different labs that have evaluated psychometric properties of the ERN (Meyer, Bress, & Proudfit, 2014; Pontifex et al., 2010). The ERN appears as a measure within the physiological unit (i.e., level) of analysis of the RDoC matrix (Weinberg et al., 2016; Weinberg, Dieterich, & Riesel, 2015) and has been proposed as a possible endophenotype for psychopathology (Manoach & Agam, 2013; Meyer, Hajcak, Torpey-Newman, Kujawa, & Klein, 2015; Olvet & Hajcak, 2008). In these ways, the ERN is a good example of a neural measure that has been robustly related to individual differences and psychopathology—and one that also is reliable in terms of psychometric properties.

Test–Retest Reliability

If a neural measure is itself trait-like, then it should be relatively consistent over testing sessions. Consistency can be considered from at least three perspectives: (a) replicability of effect, which reflects a pattern of differences between experimental conditions that replicate across samples and even participants within a sample (i.e., a within-subject comparison between conditions, such as a comparison of neural response to error vs. correct trials), (b) mean-level stability (consistent sample mean over time, such as a comparison of mean ERN for the sample between two testing occasions), and (c) rank-order stability (consistent position of a participant relative to others in the sample). Test–retest reliability refers to the latter, and is a between-subjects concept calculated as a Pearson's correlation between scores obtained at two testing occasions.

We found that the ERN has impressive test–retest reliability across two weeks (~.70; Olvet & Hajcak, 2009) and even over 2 years in both adults and children (.63 to .67; Meyer, Bress, & Proudfit, 2014; Weinberg & Hajcak, 2011). Similarly good test–retest reliability has been reported by other groups (Burwell, Malone, & Iacono, 2016; Larson, Baldwin, Good, & Fair, 2010; Segalowitz et al., 2010). The test–retest reliability of the ERN is on par with common self-report measures of individual differences (Hajcak, Huppert, Simons, & Foa, 2004), which suggests that a substantial portion of variation in the ERN is trait-like.

In the fMRI literature, there have been several studies on test–retest reliability (e.g., Bennett & Miller, 2010). As one example, fMRI-based measures of amygdala activation to fearful faces appear to have modest or poor reliability over time (Bennett & Miller, 2010; Plichta et al., 2012; Sauder, Hajcak, Angstadt, & Phan, 2013). It is important to note that a measure could have poor *test–retest* reliability *and* excellent *internal consistency*. For instance, this would be the case if individuals' amygdala activation to fearful faces was highly consistent within a scanning session but highly sensitive to between-session variability (i.e., covaried closely with current levels of anxiety, or any other differences between the scanning sessions).[2] Unfortunately, the *internal consistency* reliability of fMRI measures is almost never reported (cf. Luking et al., 2017).

The current article focuses on internal consistency reliability because this statistic has been often ignored in neuroscience research and is understood much less. Indeed, it would be good practice to report internal consistency reliability in all neuroscience articles on individual differences—particularly because it may be task- and study-dependent to some degree. This is standard for research that uses self-report data, and should be common for work using neural measures as well. It appears that neuroscience research has been lagging in terms of investigating internal con-

[2] There are instances in which one might expect test–retest reliability to be relatively low. If a measure relates to variability in states that vary from day to day, or from week to week, then test–retest reliability might decline as the duration between testing sessions increases. Additionally, test–retest reliability of neural measures *might* be lower across periods characterized by rapid neural development (e.g., childhood through adolescence). Nevertheless, it is possible that neural measures are relatively stable despite developmental changes if an individual's score from an earlier assessment predicts their later score (Meyer et al., 2014).

sistency because these calculations are less straightforward for task-based data than questionnaire data, and here we want to discuss strategies for calculating these estimates to encourage greater reporting of internal consistency.[3]

Internal Consistency Reliability

Measures of neural function reflect activity across many trials (or blocks, in the case of some fMRI designs). For what follows, it may be helpful to think about neural measures as reflecting a participant's score on a *test* that is comprised of many *items* (i.e., trials or blocks). The ERN, for instance, is quantified after averaging together all of a subjects' error trials. Averaging across many trials, like constructing a self-report scale from many items, is done to increase signal and reduce noise; doing so also increases internal consistency reliability.

Internal consistency is the most basic psychometric consideration for neural measures—and refers to the homogeneity of items on the test. In other words, if we score the ERN on every error trial, internal consistency is a measure of how similar the single-trial ERNs are across subjects. Internal consistency can be computed by scoring a neural measure on odd and even trials separately and then correlating these averages (i.e., split-half reliability, $r_{\text{odd/even}}$). Split-half reliability is a particularly practical way of quantifying internal consistency of fMRI-based measures of neural activity (Luking et al., 2017). Note that the Spearman-Brown prophecy formula is used to correct this split-half correlation because the number of items that are being considered in the overall averages is reduced by half: split-half reliability $= 2 * r_{\text{odd/even}}/1 + r_{\text{odd/even}}$.

Of course, the correlation between odd and even trials is only one way to split all items into two halves; most statistical packages will compute Cronbach's α, which is approximately the average of all possible split-half reliabilities. One distinction is that Cronbach's α requires all subjects to have the same number of trials, whereas split-half reliability (e.g., odd/even) can be computed even if participants have varying numbers of trials. In our work, these metrics of internal consistency tend to be quite similar. Across many studies, we have found that the ERN has good to excellent internal consistency, ranging from .84 to .90 (Foti, Kotov, & Hajcak, 2013; Meyer et al., 2014; Meyer, Riesel, & Proudfit, 2013; Olvet & Hajcak, 2009a, 2009b; Riesel, Weinberg, Endrass, Meyer, & Hajcak, 2013).

Task Length

In the realm of self-report measures, researchers are often concerned with maximizing reliability while minimizing test length—this is the motivation for creating short versions of longer self-report measures. But, how many items are required for a neural measure to have good internal consistency? In research on the ERN, each subject makes a variable number of errors. Thus, the ERN for one subject might be based on fewer error trials (i.e., items) than the ERN for another subject. In a series of studies, we examined the internal consistency of the ERN as more and more error trials were considered. That is, we computed internal consistency considering only the first 2 errors, then 4 errors, and so on. In these studies, we found that the internal consistency of the ERN plateaus after approximately 6 to 12 error trials (Foti et al., 2013; Meyer et al., 2013; Olvet & Hajcak, 2009b). Consistent results have been reported by other groups (Baldwin, Larson, & Clayson, 2015; Pontifex et al., 2010).

The question of *how many trials* are needed for adequate internal consistency is not specific to work on errors, where the number of items will vary across subjects. Indeed, one can ask the same question of any experimental paradigm used to study individual differences in brain activity: how many trials of each type, or blocks of trials, are sufficient to achieve adequate internal consistency? For instance, one study found that the reliability of resting-state fMRI measures increased when the scan length was increased from 5 to 13 min; another found that reward-related neural activity had comparable internal consistency when only the first half of the experiment was analyzed (Birn et al., 2013; also see Luking et al., 2017). Along the same lines then, one critical issue for clinical neuroscience studies is to optimize task length to maximize internal consistency and validity of functional neural measures. In the current study, we demonstrate how task length impacts between-groups differences in neural activity because of changes in internal consistency.

Using Neural Measures to Classify Individuals

Most between-groups studies compare neural activity in one group (i.e., generalized anxiety disorder [GAD]) to another group (i.e., healthy control participants)—and positive findings are reported in terms of statistically significant differences on the dependent variable (i.e., the ERN). Reflecting this approach, the literature is replete with variables that differ between a clinical group and a control group. However, the potential clinical utility of the neural measure is by-and-large unclear—even in the face of statistically significant differences between groups.

Ultimately, clinical neuroscience should *inform* phenotypic classification, inverting the dependent and independent variables: if an individual's ERN is known, what is the likelihood that they belong to a GAD versus healthy control group? This question pertains to how well neural measures can differentiate individuals, or diagnose healthy versus diseased states.

A simple way to examine this issue in between-groups studies is through Receiver Operating Characteristic (ROC) analyses that plot classification from the perspective of signal detection, in terms of the trade-off between sensitivity (i.e., true positive classification) and the false positive rate (i.e., 1-specificity; McFall & Treat, 1999; Swets, 1996). From an ROC plot, the success of a given cut-score can be quantified in terms of the area under the curve (AUC), which represents the probability that a randomly selected "positive" data point is higher than a randomly chosen "negative" data point.

Rather than asking whether a neural measure statistically differs between groups, ROC analyses can be used to examine *how well* a neural measure can discriminate groups—which can inform our

[3] Of note, internal consistency and test–retest reliability index somewhat different forms of error. Internal consistency indicates the level of random noise plus unique variance in items/trials (i.e., systematic variance that does not reflect the construct and is not shared by all items, e.g., fatigability of a participant may cause a drop off in neural responsivity on later trials). Test–retest reliability indicates the level of random noise plus transient effects (i.e., variance shared by all items but unique to a given occasion, e.g., anxiety during the first encounter with MRI scanner may enhance neural responsivity to fearful stimuli). Thus, both types of reliability are needed to fully describe psychometric characteristics of a test. Indeed, these indexes are not interchangeable and often are only modestly correlated (Gnambs, 2014).

thinking about whether neural measures might someday function as tests that could classify individuals. Moreover, such analyses could be useful for comparing the relative ability of two measures (e.g., neural vs. self-report) to discriminate groups. Much like reliability and between-subjects effect sizes, the ability of a neural measure to discriminate may vary with task length and trial number. Important to the purposes of the present article, both between-groups effect sizes and discriminability will depend directly on internal consistency.

The Current Study

The primary goal of the current study is to use an existing data set to demonstrate how internal consistency of neural measures might be examined to maximize their ability to index individual differences. We focus on the ERN in a sample of individuals who either met *DSM–IV* criteria for GAD, or who did not meet criteria for any *DSM–IV* disorder (i.e., healthy control participants, HC). Although the ERN is robustly related to individual differences in anxiety (Cavanagh & Shackman, 2015; Moser et al., 2013), the internal consistency of the ERN in clinically anxious individuals has only been reported in one study to date (Baldwin et al., 2015). First, we examine the ERN in both GAD and HC participants as a function of increasing errors—to examine the relationship between task length (i.e., number of items) and between-groups effect sizes. Second, we examine internal consistency reliability of the ERN in both GAD and HC participants—both overall using split-half (i.e., odd vs. even) reliability and Cronbach's α, and Cronbach's α as a function of increasing errors. We show that internal consistency is directly related to between-groups effect sizes. Finally, we report ROC analyses overall and as a function of increasing errors to assess the ability of the ERN to discriminate GAD from HC as a function of task length. The broader goal of the current study is to stimulate thinking and empirical examination of common neural and psychophysiological metrics in relation to psychometric properties—and how we might apply these analyses in other data sets to construct more efficient and *reliable* neural individual difference measures.

Method

Participants

The current study combines samples from two separate previously published studies that examined ERN in relation to GAD (Weinberg, Klein, & Hajcak, 2012; Weinberg, Olvet, & Hajcak, 2010; the current sample was also reported on in Meyer, Lerner, Reyes, Laird, & Hajcak, 2017). All research was approved by the Stony Brook University Institutional Review Board. The current study focuses on 41 participants with a diagnosis of GAD (but not comorbid depression), and 53 individuals with no current *DSM* diagnosis (i.e., healthy controls, HC). All diagnoses were made using the Structured Clinical Interview for *DSM-IV* (SCID). For additional information on recruiting and patient information, see Weinberg et al., 2010, 2012).

Task and Materials

An arrow version of the flanker task (Eriksen & Eriksen, 1974) was administered on a Pentium D class computer, using Presentation software (Neurobehavioral Systems, Inc., Albany, CA) to control the presentation and timing of all stimuli. Each stimulus was displayed on a 19 in (48.3 cm) monitor. On each trial, five horizontally aligned arrowheads were presented. Half of all trials were compatible ("< < < < <" or "> > > > >") and half were incompatible ("< < > < <" or "> > < > >"). The order of compatible and incompatible trials was random. Each set of arrowheads occupied approximately 1.3° of visual angle vertically and 9.2° horizontally. All stimuli were presented for 200 ms followed by an intertribal interval (ITI) that varied randomly from 2,300 to 2,800 ms.

Procedure

After informed consent and a brief description of the experiment, EEG electrodes were attached and the subject was given detailed task instructions. All participants performed multiple tasks during the experiment. The order of the tasks was counterbalanced across participants and the results of other tasks will be reported elsewhere. Participants were seated facing a computer screen at a viewing distance of approximately 24 in (61 cm) and were instructed to press the right mouse button if the center arrow was facing to the right and to press the left mouse button if the center arrow was facing to the left. Information about each response (e.g., RT, accuracy), was recorded. Participants performed a practice block containing 30 trials during which they were instructed to respond both as accurately and quickly as possible. The actual task consisted of 11 blocks of 30 trials (330 trials total) with each block initiated by the participant. Participants received feedback based on their performance at the end of each block. If performance was 75% correct or lower, the message "Please try to be more accurate" was displayed. Performance above 90% correct was followed by "Please try to respond faster." If performance was between 75 and 90% correct, the message "You're doing a great job" was displayed.

Psychophysiological Recording, Data Reduction, and Analysis

Continuous EEG recordings were collected using an elastic cap and the ActiveTwo BioSemi system (BioSemi, Amsterdam, Netherlands). Thirty-four electrode sites were used, based on the 10/20 system, as well as two electrodes on the right and left mastoids. The electrooculogram (EOG) generated from eye movements and eyeblinks was recorded using four facial electrodes: horizontal eye movements (HEM) were measured via two electrodes located approximately 1 cm outside the outer edge of the right and left eyes. Vertical eye movements (VEM) and blinks were measured via two electrodes placed approximately 1 cm above and below the right eye. The EEG signal was preamplified at the electrode to improve the signal-to-noise ratio and was digitized at 24-bit resolution with a LSB value of 31.25 nV and a sampling rate of 1024 Hz, using a low-pass fifth order sinc filter with −3dB cutoff point at 208 Hz. Each active electrode was measured online with respect to a common mode sense (CMS) active electrode, located between PO3 and POz, producing a monopolar (nondifferential) channel. CMS forms a feedback loop with a paired driven right leg (DRL) electrode. Offline, all data were referenced to the average of the left and right mastoids, and band-pass filtered with low and high cutoffs of 0.1 and 30 Hz, respectively. Eyeblink and ocular corrections were conducted using both VEM and HEM channels per a modification of the original algorithm published in Gratton, Coles, and Donchin (1983).

A semiautomatic procedure was used to detect and reject artifacts. Data from individual channels were rejected if a voltage step

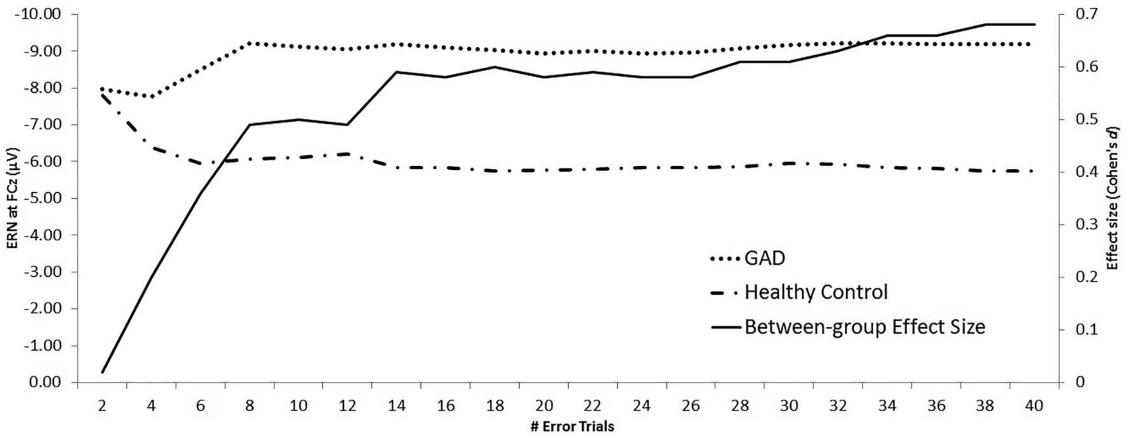

Figure 1. Error-related negativity (ERN) amplitude (left ordinate) for GAD (dotted line) and Healthy Control (dashed/dotted line) participants, and between-groups effect size (right ordinate; solid line), as a function of increasing error trials. GAD = generalized anxiety disorder.

of more than 50.0 μV between sample points or a voltage difference of 300.0 μV within a trial existed. In addition, data were identified as artifactual if a voltage difference of less than .50 μV within 100 ms intervals was present. Visual inspection of the data was then conducted to detect and reject any remaining artifacts.

The EEG signal was segmented for each trial beginning 500 ms before error response onset and continuing for 1,500 ms (i.e., 1,000 ms after the response); a 200 ms window from −500 to −300 ms before the response onset served as the baseline. The ERN was scored on each trial as the average activity from 0 to 100 ms at FCz, after error responses. From these values, we computed ERN averages as increasing number of errors were analyzed (i.e., the first 2 errors, the first 4 errors, etc.). These values were then used for computing metrics of internal consistency, effect size, and discriminability for participants in the GAD versus HC group. In addition, the ERN was also computed for odd and even trials separately for split-half reliability analyses.

Results

Between-Groups Effect Size

In line with our previous reports from these data, individuals with GAD were characterized by a larger overall (i.e., grand average) ERN ($M = -8.92$, $SD = 4.49$) than HC participants ($M = -5.61$, $SD = 5.17$, $t(92) = 3.25$, $p < .01$), and this was associated with a medium to large effect size (Cohen's $d = .68$). Figure 1 presents the average ERN for GAD and HC as a function of increasing trial numbers. A more negative ERN emerged in GAD within the first few trials and the magnitude of this difference was quite stable as more error trials were included in analyses. Figure 1 also presents the effect size (in Cohen's *d*) for between-Group ERN comparisons as increasing error trials were examined. Group difference increased dramatically from 2 to 8 trials (i.e., Cohen's *d* of .48), further improvement was less dramatic but the difference increased to $d = .59$ by 14 trials; increases in trial number after 14 drove group difference even higher, but the gains were subtle, with *d* increasing only to .68 by 38 trials.

Internal Consistency

The Spearman-Brown corrected split-half reliability of the ERN was .71 for the GAD and .75 for the HC group. Figure 2 (top) presents Cronbach's α for ERN within the GAD and HC groups as increasing errors were examined. Cronbach's α was first examined considering only the first two error trials, then the first four error trials, and so on. For example, if a subject had 15 errors, then they would have been included in all analyses up to 14 errors. Figure 2 (bottom) presents the percentage of participants per group that were included in each average. In both GAD and HC, the increase in reliability followed the same pattern as increases in effect size (i.e., large gains through 8 trials, modest increase through 14 trials, and subtle increases thereafter). Cronbach's α reached a maximum of about .75 to .85, which corresponded closely to the split-half values. In addition, as more trials were examined, Cronbach's α increased only slightly—and it is important to note that there was significant loss of subjects as more error trials were included in these analyses.

ROC Analyses

Figure 3 presents the ROC curve for the overall grand averaged ERN (AUC = .69 $p < .002$ [95% confidence interval, CI: .58–.80]). Figure 4 plots AUC values as a function of increasing error trials.[4] The ability to discriminate GAD from HC peaked at 8 error trials and effectively plateaued after.

Internal Consistency and Between-Groups Effects

To further highlight the relationship between internal consistency and between-groups measures of effect size and discrimination, Figure 5 (top) presents the scatterplot between Cronbach's α and Cohen's

[4] To facilitate comparisons with criterion correlations that have been reported for other physiological measures, the point-biserial correlation between ERN and GAD status as a function of increasing error trials is: .009, .096, .177, .235, .241, .238, .282, .278, .286, .278, .281, .276, .277, .291, .288, .299, .312, .310, .319, and .318.

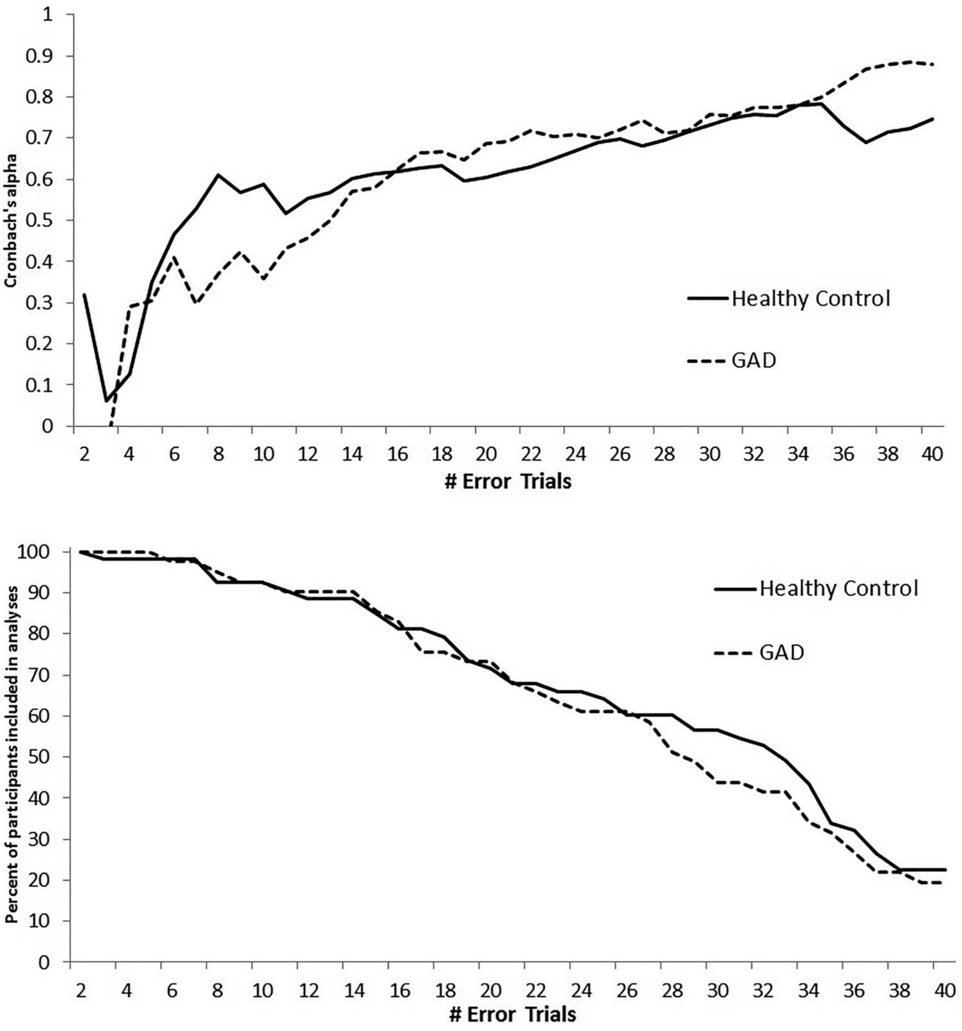

Figure 2. Cronbach's α for GAD (solid) and Healthy Control (dashed) participants as a function of increasing error trials (top) and percentage of participants included in these analyses (bottom). GAD = generalized anxiety disorder.

d for ERN values from Figures 2 and 1, respectively. Each data point represents the internal consistency of the ERN calculated after a specific number of error trials (i.e., 2, 4, 6 . . . 40) and the between-groups effect size based on that ERN. Similarly, Figure 5 (bottom) plots internal consistency against AUC values using data from Figures 2 and 4, respectively. For both plots, internal consistency was corrected by taking the square root of the product of Cronbach's α for GAD and Healthy Control subjects. As evident from Figure 5, internal consistency is highly correlated with both between-groups effect size, $r = .94$, $p < .001$ and AUC, $r = .83$, $p < 001$.

Reanalyses With Subjects Who Made More Than 20 Errors

As is evident from Figure 2 (bottom), one confound in the preceding analyses is that fewer participants were included in analyses that focused on more errors. However, subjects were not excluded randomly: analyses focusing on 34 errors would necessarily exclude more accurate subjects who made relatively few errors. To rule-out this possible confound, we reran all analyses examining overall split-half reliability (.81 in GAD, .77 in HC), as well as the impact of number of error trials (up to 20) on coefficient α (Figure 6, top), between-groups effect size (Figure 6, bottom), and ROC analyses (Figure 6, bottom), focusing only on the 31 GAD and 42 HC participants who made at least 20 errors. These results were quite consistent with the overall analyses: internal consistency, between-groups effect size, and the ability of the ERN to discriminate GAD from HC increased dramatically up to 8 trials, and increased more modestly as more error trials were examined.

Discussion

In the current study, both GAD and HC had comparable internal consistency of the ERN—both overall (i.e., split-half) and as a function of increasing errors. These data are consistent with other

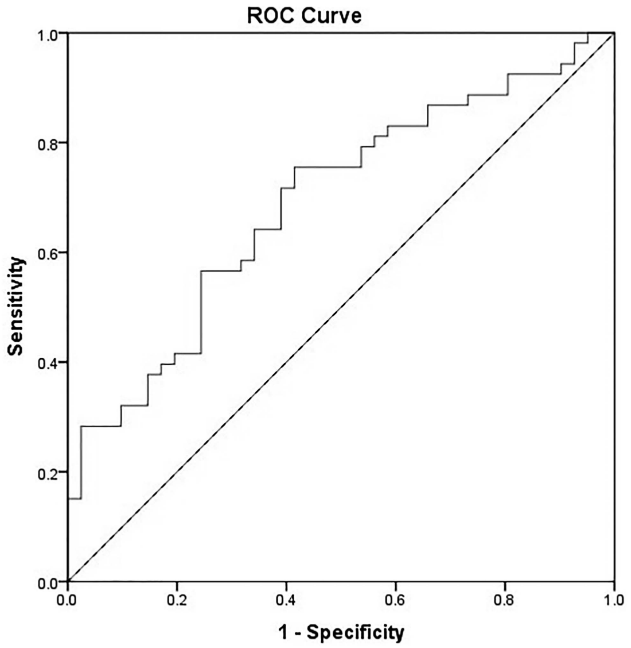

Figure 3. Receiver operator characteristic (ROC) curve illustrating performance of overall ERN in classifying GAD and Healthy Control participants (solid line); sensitivity is plotted as a function of the false positive rate (i.e., 1-sensitivity); the area under the curve (AUC) is .69, and chance classification (.50 AUC) is plotted for comparison (dashed line). GAD = generalized anxiety disorder.

research studies that have reported high internal consistency reliability of the ERN in various forms of psychopathology (Baldwin et al., 2015; Foti et al., 2013). Moreover, the internal consistency data also informed the interpretation of between-subjects comparisons. Specifically, the between-subjects effect size was large after eight trials, and was maximal and effectively plateaued after about 14 errors. Similarly, the ability of the ERN to discriminate GAD from HC—quantified in terms of AUC using ROC analyses—reached its maximum and was stable after approximately eight error trials. Indeed, both between-subjects effects were highly correlated with internal consistency. It is important to note that similar results were obtained when we reanalyzed only subjects who committed at least 20 errors. Overall, these data demonstrate how between-groups effects and discriminability depend on internal consistency—the pattern of increasing internal consistency was the basis for similarly increasing Cohen’s *d* and AUC.

It is important to note that psychometric properties of neural measures are not fixed; rather, they will vary across populations, tasks, and labs—and it would be sensible to report internal consistency of neural measures in every publication and data set. As an example, we measured the ERN from the same subjects who performed three different tasks; although errors in each task elicited a similar *looking* ERN, the correlation between these ERNs was modest—and the internal consistency varied substantially across tasks (Meyer et al., 2014, 2013; Riesel et al., 2013). Thus, ostensibly similar neural responses observed in two apparently comparable tasks may not correlate highly with one another, and one cannot assume that neural measures from different tasks have equivalent psychometric properties.

The current data further suggest concrete ways in which psychometric analyses can be applied to optimize task design. The increased neural response to errors in GAD did not increase or decrease across the testing session in the current data. After approximately 10 errors, more errors did not lead to appreciable increases in reliability, between-groups effect sizes, or better between-groups discriminability. These data suggest that the clinical utility of the ERN may be maximized by relatively shorter tasks. In the case of the ERN, it may be more practical to utilize an adaptive task that varies in length across individuals, where task length would be determined by error counts rather than having the same total number of trials across individuals. It is important to consider the possibility that task length may impact between-groups differences and effect sizes (i.e., the amount of variation of a measurement that relates to another individual difference measure). Suppose, for instance, that individual differences in a neural measure were large early during a task and became smaller later in the task; alternatively, individual differences may emerge and become larger over the course of a task. In either case, it would be

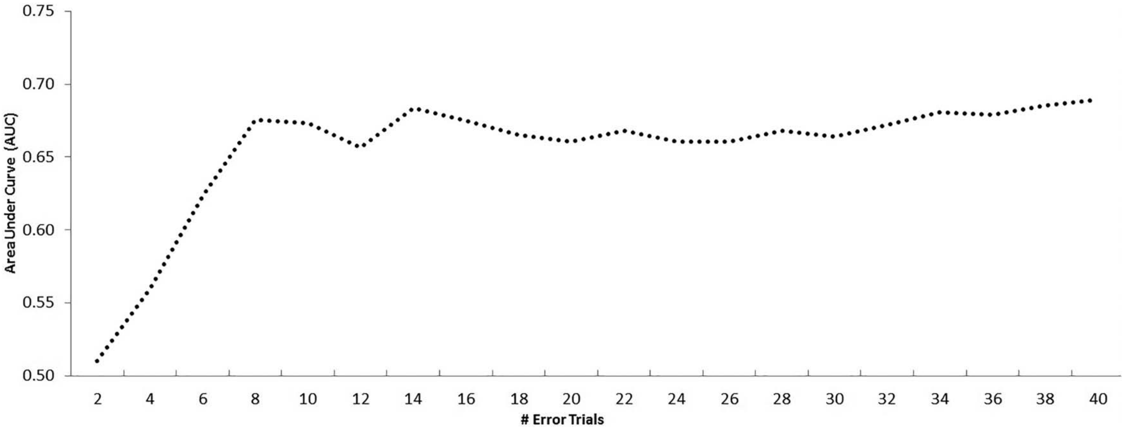

Figure 4. Area under the curve (AUC) for receiver operator characteristic (ROC) analyses performed as a function of increasing error trials.

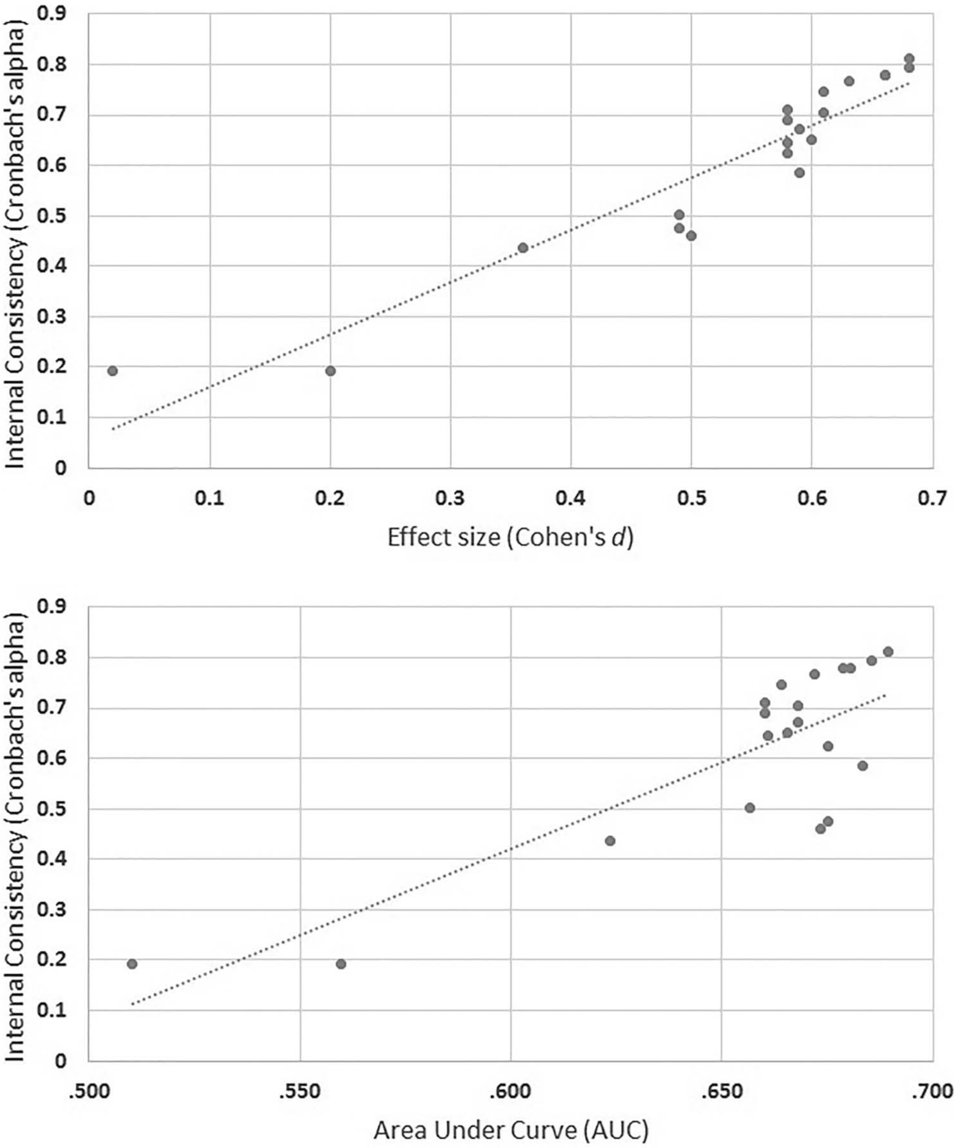

Figure 5. Internal consistency (corrected Cronbach's α) plotted against: effect size (Cohen's *d*; top) and area under curve (AUC; bottom).

necessary to examine the impact of increasing trials on the magnitude of cross-variation in scores on the neural measure of interest (e.g., between-groups effect sizes) to evaluate this possibility empirically. In evaluating task length and number of trials, clinical neuroscience studies need to balance both reliability and between-subjects effect sizes. An optimal task for clinical utility would be the shortest possible to maximize both reliability and between-subjects differences.

The current study demonstrates how, within a task, internal consistency can limit between-groups effects. Reliability limits criterion validity (i.e., the degree to which a neural indicator relates to another individual difference variable). It is important to note, however, that higher reliability does not automatically translate to increased validity. Individual differences in anxiety, for instance, could be more highly correlated with a measure that has lower internal consistency than the ERN if more of that measure's true score variance relates to anxiety. As a practical example, consider difference scores. In many clinical neuroscience studies, the dependent variable reflects a difference between two conditions; that is, many studies are interested in individual differences *of a condition effect, quantified as a difference score*. In the preceding discussion, we focused on the ERN. Many studies have examined individual differences in anxiety in relation to neural activity that differentiates error from correct trials (i.e., the difference between the ERN and the correct response negativity, or CRN; Moser et al., 2013). Although a full discussion of subtraction-based difference scores is beyond the scope of the current article (Meyer et al., 2017), the logic here is to examine brain activity on

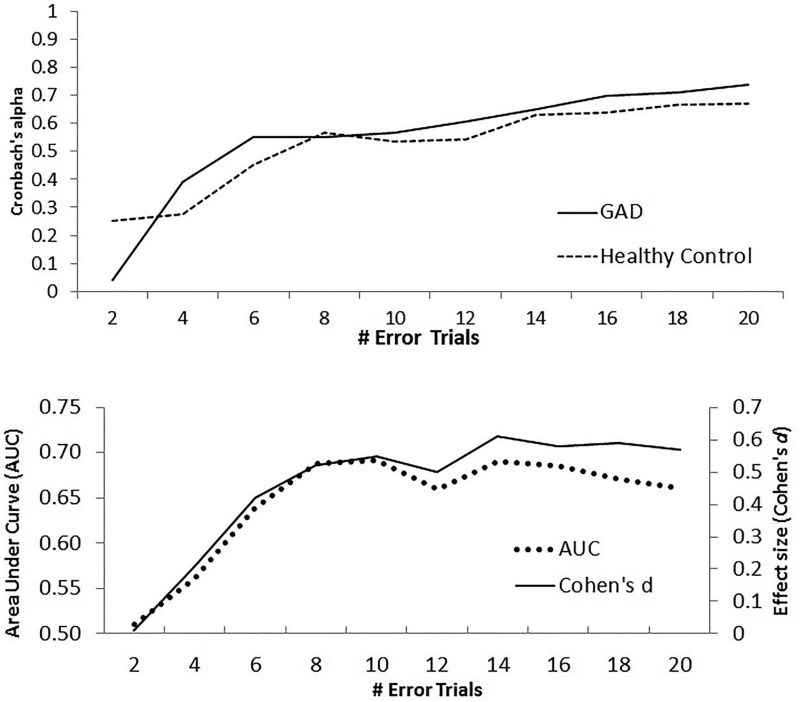

Figure 6. Among participants who made at least 20 errors, Cronbach's α is plotted for GAD (N = 31; solid) and Healthy Control (N = 42; dashed) participants as a function of increasing error trials (top); also plotted are area under the curve (AUC; left ordinate) and effect size (Cohen's d; right ordinate) for ROC and between-groups comparisons, respectively, as a function of increasing error trials (bottom). GAD = generalized anxiety disorder.

error trials relative to what is observed on correct trials—to isolate brain activity that is *specific* to errors (i.e., the ΔERN, or ERN minus CRN). Along similar lines, fMRI studies commonly use subtraction-based analyses.

In the context of the current article, difference scores tend to have lower reliability (Luking et al., 2017). For instance, whereas the internal consistency of the ERN (.84–.88) and CRN (.98) are both excellent, the internal consistency of the ΔERN in these same data is .67–.76 (Olvet & Hajcak, 2009a, 2009b). Essentially this is because the ERN and CRN are highly correlated with one another. The reliability of a difference score (C, where C = A - B) will depend on the reliability of A, the reliability of B, and the correlation between A and B. The reliability of a difference score (i.e., C) will always be less reliable than the constituent scores (i.e., A and B)—and the higher the correlation between A and B, the lower the reliability of the difference score, C.

Yet, anxiety seems about equally correlated with the ERN and ΔERN (Moser et al., 2013). In other words, the criterion validity of the ERN and ΔERN are similar, despite the former having better psychometric properties; this is presumably because ΔERN contains the same or more true score variance that relates to anxiety. More broadly, it is possible that a measure with relatively modest internal consistency could have adequate criterion validity—though the magnitude of the latter will be limited by the former.

Although the internal consistency, between-groups effect size, and the ability of the ERN to discriminate GAD from HC is on par with self-report measures, we would note that the ERN is only one neural metric that may be relevant for anxiety and related disorders. Indeed, the level of observed discrimination of diagnostic status and between-groups effect sizes based on the ERN alone suggests a need for improved criterion validity. Rather than conceptualizing the ERN as a test, it might be more informative to think about the ERN as a single item on a broader composite biomarker scale (e.g., Patrick & Hajcak, 2016). Of course, that approach requires examining how measures like the ERN relate to other biological measures of individual differences—this is an issue of construct validity (Cronbach & Meehl, 1955). As one example, we have found that individual differences in the ERN relate to increased startle potentiation to unpleasant images (Meyer et al., 2017)—findings that suggest conceptual overlap between these measures. By aggregating across multiple measures, it would be feasible to do factor analyses on biomarkers to examine "subscales" of measures that may hang together (Nelson, Patrick, & Barnat, 2011; Patrick & Bernat, 2010), and in relation to other individual difference constructs indexed by self-report measures.

Neuroscience has shed significant light on how the brain functions—and it is a sensible organ to study to better understand individual differences in normative and abnormal behavior. Yet, neuroscientific measures have not led to significant changes in the way we diagnose, treat, or prevent mental illness. From the perspective of the RDoC initiative, this failure may be an inevitable consequence of a diagnostic system with poor validity. A further obstacle may be that clinical neuroscience progress has been limited because the field has used *robust* neural measures with unknown or low reliability. Clinical neuroscience has concerned itself almost exclusively with criterion validity—whether neural measures correlate with individual differences.

The point of the current article is that internal consistency is *another* important consideration for interpreting neural measures in terms of individual differences. Indeed, a significant group difference or between-subjects correlation represents a weak finding in the absence of psychometric information. The internal consistency of two measures places an upper limit on their possible correlation: if two measures have internal consistencies of .70 and .60, their maximum *possible* correlation is r = .64; anything higher would likely fail to replicate and reflect Type I error.[5] The potential impact of unreliable measures needs to be taken seriously, and scientific articles should be required to report internal consistency of neural measures as a prerequisite for examining individual differences. Routine reporting of internal consistency in neuroscientific studies would be an important step toward improving neuroscientific research on individual differences. Reporting internal consistency could help the neuroscience field better interpret, and possibly improve, between-subjects effects—as well as potentially reduce failures to replicate. This is standard practice in research using self-report measures. We should approach measures with poor internal consistency with caution (e.g., Rodebaugh et al., 2016). Poor internal consistency reduces statistical power and the utility of an individual difference measure. The current article demonstrates how internal consistency constrains between-subjects effects. The specific analyses presented here could be done with many neural measures in regard to psychometric properties and individual differences research.

[5] Validity is constrained by the square root of the product of the reliability coefficients.

References

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

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Year in Cognitive Neuroscience 2010,* 1191, 133–155. http://dx.doi.org/10.1111/j.1749-6632.2010.05446.x

Birn, R. M., Molloy, E. K., Patriat, R., Parker, T., Meier, T. B., Kirk, G. R., . . . Prabhakaran, V. (2013). The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage, 83,* 550–558. http://dx.doi.org/10.1016/j.neuroimage.2013.05.099

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. http://dx.doi.org/10.1111/psyp.12646

Cavanagh, J. F. J., & Shackman, A. J. (2015). Frontal midline theta reflects anxiety and cognitive control: Meta-analytic evidence. *Journal of Physiology, 109,* 3–15. http://dx.doi.org/10.1016/j.jphysparis.2014.04.003

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

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine, 11,* 126. http://dx.doi.org/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. http://dx.doi.org/10.1037/a0034028

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon identification of a target letter in a non- search task. *Perception & Psychophysics, 16,* 143–149. http://dx.doi.org/10.3758/BF03203267

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

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52,* 20–28. http://dx.doi.org/10.1016/j.jrp.2014.06.003

Gottesman, I. I., & Gould, T. D. (2003). The endophenotype concept in psychiatry: Etymology and strategic intentions. *The American Journal of Psychiatry, 160,* 636–645. http://dx.doi.org/10.1176/appi.ajp.160.4.636

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

Hajcak, G., Huppert, J. D., Simons, R. F., & Foa, E. B. (2004). Psychometric properties of the OCI-R in a college sample. *Behaviour Research and Therapy, 42,* 115–123. http://dx.doi.org/10.1016/j.brat.2003.08.002

Hajcak, G., & Patrick, C. J. (2015). Situating psychophysiological science within the Research Domain Criteria (RDoC) framework. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology, 98,* 223–226. http://dx.doi.org/10.1016/j.ijpsycho.2015.11.001

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *The American Journal of Psychiatry, 167,* 748–751. http://dx.doi.org/10.1176/appi.ajp.2010.09091379

Kendler, K. S., & Neale, M. C. (2010). Endophenotype: A conceptual analysis. *Molecular Psychiatry, 15,* 789–797. http://dx.doi.org/10.1038/mp.2010.8

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., . . . Kendler, K. S. (1994). Lifetime and 12-month prevalence of *DSM–III–R* psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry, 51,* 8–19. http://dx.doi.org/10.1001/archpsyc.1994.03950010008002

Kozak, M. J., & Cuthbert, B. N. (2016). The NIMH research domain criteria initiative: Background, issues, and pragmatics. *Psychophysiology, 53,* 286–297. http://dx.doi.org/10.1111/psyp.12518

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.

Lilienfeld, S. O. S. (2014). The Research Domain Criteria (RDoC): An analysis of methodological and conceptual challenges. *Behaviour Research and Therapy, 62,* 129–139. http://dx.doi.org/10.1016/j.brat.2014.07.019

Luking, K. R., Nelson, B. D., Infantolino, Z. P., Sauder, C. L., & Hajcak, G. (2017). Internal consistency of fMRI and EEG measures of reward in late childhood and early adolescence. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 2,* 289–297.

Manoach, D. S., & Agam, Y. (2013). Neural markers of errors as endophenotypes in neuropsychiatric disorders. *Frontiers in Human Neuroscience, 7,* 350.

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241. http://dx.doi.org/10.1146/annurev.psych.50.1.215

Meehl, P. E. (1986). Diagnostic taxa as open concepts: Metatheoretical and statistical questions about reliability and construct validity in the grand strategy of nosological revision. In T. Million & G. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 215–231). New York, NY: Guilford.

Meehl, P. E. (1995). Bootstraps taxometrics. Solving the classification problem in psychopathology. *American Psychologist, 50,* 266–275. http://dx.doi.org/10.1037/0003-066X.50.4.266

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

Meyer, A., Hajcak, G., Glenn, C. R., Kujawa, A. J., & Klein, D. N. (2017). Error-related brain activity is related to aversive potentiation of the startle response in children, but only the ERN is associated with anxiety disorders. *Emotion, 17,* 487–496. http://dx.doi.org/10.1037/emo0000243

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

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

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

Miller, G. A. (2010). Mistreating psychology in the decades of the brain. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 5,* 716–743. http://dx.doi.org/10.1177/1745691610388774

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

Nelson, L. D., Patrick, C. J., & Bernat, E. M. (2011). Operationalizing proneness to externalizing psychopathology as a multivariate psychophysiological phenotype. *Psychophysiology, 48,* 64–72. http://dx.doi.org/10.1111/j.1469-8986.2010.01047.x

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

Olvet, D. M., & Hajcak, G. (2009a). Reliability of error-related brain activity. *Brain Research, 1284,* 89–99. http://dx.doi.org/10.1016/j.brainres.2009.05.079

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

Patrick, C. J., & Bernat, E. M. (2010). Neuroscientific foundations of psychopathology. In T. Millon, R. F. Krueger, & E. Simonsen (Eds.), *Contemporary directions in psychopathology: Scientific foundations of the DSM-V and ICD-11* (pp. 419–452). New York, NY: Guilford Press.

Patrick, C. J., & Hajcak, G. (2016). RDoC: Translating promise into progress. *Psychophysiology, 53,* 415–424. http://dx.doi.org/10.1111/psyp.12612

Phan, K. L., Wager, T., Taylor, S. F., & Liberzon, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage, 16,* 331–348. http://dx.doi.org/10.1006/nimg.2002.1087

Plichta, M. M., Schwarz, A. J., Grimm, O., Morgen, K., Mier, D., Haddad, L., . . . Meyer-Lindenberg, A. (2012). Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *NeuroImage, 60,* 1746–1758. http://dx.doi.org/10.1016/j.neuroimage.2012.01.129

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.

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. http://dx.doi.org/10.1016/j.biopsycho.2013.04.007

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., . . . Lenze, A. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology, 125,* 840–851. http://dx.doi.org/10.1037/abn0000184

Sanislow, C. A., Pine, D. S., Quinn, K. J., Kozak, M. J., Garvey, M. A., Heinssen, R. K., . . . Cuthbert, B. N. (2010). Developing constructs for psychopathology research: Research domain criteria. *Journal of Abnormal Psychology, 119,* 631–639. http://dx.doi.org/10.1037/a0020909

Sauder, C. L., Hajcak, G., Angstadt, M., & Phan, K. L. (2013). Test-retest reliability of amygdala response to emotional faces. *Psychophysiology, 50,* 1147–1156. http://dx.doi.org/10.1111/psyp.12129

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. http://dx.doi.org/10.1111/j.1469-8986.2009.00942.x

Stringaris, A. (2015). Editorial: Neuroimaging in clinical psychiatry—When will the pay off begin? *Journal of Child Psychology and Psychiatry, 56,* 1263–1265. http://dx.doi.org/10.1111/jcpp.12490

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tovote, P., Fadok, J. P., & Lüthi, A. (2015). Neuronal circuits for fear and anxiety. *Nature Reviews Neuroscience, 16,* 317–331. http://dx.doi.org/10.1038/nrn3945

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

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

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

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

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

Received August 9, 2016
Revision received March 22, 2017
Accepted March 23, 2017 ■