Psychophysiology, 53 (2016), 1241–1255. Wiley Periodicals, Inc. Printed in the USA.

DOI: 10.1111/psyp.12663

Psychometric properties of startle and corrugator response in NPU, affective picture viewing, and resting state tasks

JESSE T. KAYE, DANIEL E. BRADFORD, AND JOHN J. CURTIN
Department of Psychology, University of Wisconsin–Madison, Madison, Wisconsin, USA

Abstract

The current study provides a comprehensive evaluation of critical psychometric properties of commonly used psychophysiology laboratory tasks/measures within the NIMH RDoC. Participants ($N = 128$) completed the no-shock, predictable shock, unpredictable shock (NPU) task, affective picture viewing task, and resting state task at two study visits separated by 1 week. We examined potentiation/modulation scores in NPU (predictable or unpredictable shock vs. no-shock) and affective picture viewing tasks (pleasant or unpleasant vs. neutral pictures) for startle and corrugator responses with two commonly used quantification methods. We quantified startle potentiation/modulation scores with raw and standardized responses. We quantified corrugator potentiation/modulation in the time and frequency domains. We quantified general startle reactivity in the resting state task as the mean raw startle response during the task. For these three tasks, two measures, and two quantification methods, we evaluated effect size robustness and stability, internal consistency (i.e., split-half reliability), and 1-week temporal stability. The psychometric properties of startle potentiation in the NPU task were good, but concerns were noted for corrugator potentiation in this task. Some concerns also were noted for the psychometric properties of both startle and corrugator modulation in the affective picture viewing task, in particular, for pleasant picture modulation. Psychometric properties of general startle reactivity in the resting state task were good. Some salient differences in the psychometric properties of the NPU and affective picture viewing tasks were observed within and across quantification methods.

Descriptors: Analysis/statistical methods, Startle blink, EMG, Emotion, Anxiety, Stress

Psychophysiological tasks are poised to become a major contributor to the National Institute of Mental Health (NIMH) Research Domain Criteria (RDoC) and related initiatives in experimental medicine (Hajcak & Patrick, 2015; Insel, 2015; Patrick & Hajcak, 2016). The RDoC provides a novel framework to examine "basic dimensions of functioning underlying the full range of human behavior from normal to abnormal" across multiple levels of analysis (NIMH, 2015). Psychophysiological tasks are attractively situated to index dimensional individual differences relevant to a wide array of applications within and beyond the RDoC initiative. For instance, they may tap RDoC domains (e.g., negative valence systems, positive valence systems) at a level of analysis that can bridge critically between lower (e.g., neural circuits) and higher (e.g., behavior, self-reports) levels. However, these tasks are by no means only valuable within the context of the RDoC initiative. They also align well with NIMH's current research priorities in experimental medicine including a salient attention to mechanisms in clinical trials (Insel, 2015; Insel & Gogtay, 2014), the use of surrogate end points in FAST fail initiatives for treatment development (Insel, 2015), and endophenotype identification (Miller & Rockstroh, 2013), to name just a few applications. Equally important, the proliferation of "turn-key" systems has made the implementation of these tasks more efficient and affordable for scientists without extensive specialized training or large research grants. However, for psychophysiological tasks to meaningfully contribute to the study of individual differences, whether in the context of the goals of the RDoC or broader applications, they must possess sound psychometric properties. Are they up for the task?

Prominent psychologists have recently issued strong calls to validate critical psychometric properties of laboratory tasks including their stability, internal consistency, and robustness (Cuthbert, 2014; Hajcak & Patrick, 2015; Lilienfeld, 2014). Psychology and related disciplines have a rich history of rigorous psychometric evaluation of their self-report measures, but the reliability and validity of physiological measures within laboratory tasks have often been presumed rather than demonstrated (Lilienfeld, 2014). In the face of mounting public and academic concern about the reliability of psychological science broadly, the research described here and in other contributions to this special issue of *Psychophysiology* represents important steps to answer this call for rigorous,

Funding was provided by the National Science Foundation (DGE-0718123), National Institute of Alcohol Abuse and Alcoholism (F31 AA022845), and the National Institute of Drug Abuse (R01 DA033809). We would like to thank Charles Rohrer, Austin Kayser, Rachel Korhumel, and Rachel Hamilton for their assistance with data collection.

Address correspondence to: Jesse T. Kaye, Department of Psychology, University of Wisconsin–Madison, 1202 West Johnson St., Madison, WI, 53706, USA. E-mail: jtkaye@wisc.edu

comprehensive, and systematic evaluation of the psychometric properties of our psychophysiological tasks.

Psychophysiological Tasks in the RDoC Framework

We designed the current study to evaluate key psychometric properties of three commonly used psychophysiological tasks that can be anchored within the RDoC: (1) the no-shock, predictable shock, unpredictable shock task (NPU task), (2) the affective picture viewing task, and (3) the resting state task.

The NPU task (Schmitz & Grillon, 2012) manipulates participants' affect by administering mild electric shocks that are predictable in one condition and unpredictable in another condition. The contrast of physiological responding selectively during predictable or unpredictable shock conditions with a neutral, no-shock condition has been proposed to map onto acute and potential threat constructs, respectively, within the RDoC negative valence system. The NPU task has been used extensively to study mood and anxiety disorders (Grillon et al., 2008, 2009), addiction (Bradford, Curtin, & Piper, 2015; Hogle, Kaye, & Curtin, 2010), and the effects of pharmaceutical and recreational drugs (Grillon et al., 2006; Moberg & Curtin, 2009).

The affective picture viewing task manipulates participants' affect with unpleasant, pleasant, and neutral pictures while measuring their concomitant physiological responses. Lang, Bradley and Cuthbert (1993, 2008) developed the International Affective Picture System (IAPS) to provide a standardized set of unpleasant, pleasant, and neutral pictures for use within this task. The contrast of physiology during unpleasant versus neutral pictures is situated clearly within the negative valence system domain, although the specific construct has not been precisely specified to date. Additionally, the contrast of pleasant versus neutral pictures can contribute to measurement of the positive valence system domain. Over the past 2 1/2 decades, this task has been used to examine a wide array of topics related to affective, cognitive, social, and clinical science (Bradley & Lang, 2007).

The resting state task, often conducted at "baseline" prior to the start of another focal experimental task, involves the measurement of physiology during a period of time characterized by the absence of other explicit manipulations or potent experimental stimuli. This task can serve as a covariate in the analysis of physiology during those focal tasks to increase power and precision (Bradford, Kaye, & Curtin, 2014). However, physiology in the resting state task may also reflect important individual differences in traitlike, dispositional functioning relevant to a variety of RDoC domains. For example, general startle reactivity during this resting state task has been suggested to index individual differences in defensive reactivity (Bradford, Kaye, & Curtin, 2014; Vaidyanathan, Patrick, & Cuthbert, 2009). It may also be a useful index of valence-neutral arousal, which may have relevance to the RDoC matrix.

The acoustic startle response and corrugator response are both commonly measured within each of these three tasks and serve as the focal physiological measures in the current study, given their well-validated connections to affective response. The startle response is a defensive reflex elicited by brief, startling acoustic noise probes and measured via electromyography (EMG) activity in the orbicularis oculi muscle associated with the human eyeblink startle reflex. The startle response is consistently potentiated under conditions of acute or potential threat and occasionally attenuated during presentation of positively valenced stimuli (e.g., pleasant pictures; Bradley, Codispoti, Cuthbert, & Lang, 2001). The corrugator response is measured via EMG activity in the corrugator muscle associated with the human facial frown. As with the startle response, it is also bidirectionally modulated by the valence of foreground stimuli such that it is increased during unpleasant and occasionally decreased during pleasant stimuli (Larsen, Norris, & Cacioppo, 2003). These physiological measures are specifically highlighted in the RDoC framework and represent prime candidates to tap the negative valence system domain constructs (and additionally the positive valence system domain selectively in the affective picture viewing task) when measured within these three tasks. Startle and corrugator consistently display robust responses to threatening stimuli. Furthermore, unlike some other peripheral physiology measures (e.g., heart rate, skin conductance), they are not driven primarily by arousal but are differentially sensitive to negative and positive valence. Not surprisingly, researchers have begun to examine their psychometric properties within these tasks, as we describe next.

Psychometric Properties of Psychophysiological Tasks

To date, only Shankman and colleagues (2013) have evaluated the psychometric properties of predictable and unpredictable shock (vs. no-shock) startle potentiation in the NPU task. They reported that the temporal stability of startle potentiation was adequate (rs = ~.70) for both predictable and unpredictable shock startle potentiation. The effect size stability and internal consistency of predictable and unpredictable shock startle potentiation have not yet been reported.[1] Furthermore, no psychometric properties of corrugator potentiation in the NPU task have been reported.

More attention has been paid to the psychometric properties of unpleasant and pleasant (vs. neutral) picture modulation in the affective picture viewing task, but troubling inconsistencies have been reported and key gaps in our knowledge remain. Reports of the temporal stability of unpleasant and pleasant picture startle modulation has varied widely across studies (correlations range from 0.16 to 0.55 for unpleasant modulation and −0.06 to 0.44 for pleasant modulation; Larson, Ruffalo, Nietert, & Davidson, 2000, 2005; Lee, Shackman, Jackson, & Davidson, 2009; Manber, Allen, Burton, & Kaszniak, 2000). These inconsistent and at times very low correlations have led some to call into question the reliability of startle modulation in the affective picture viewing task (Heller, Greischar, Honor, Anderle, & Davidson, 2011; Lee et al., 2009). In contrast, the temporal stability of unpleasant picture modulation for corrugator has been observed to be consistently adequate or better (correlations: .61 to .84; Lee et al., 2009; Manber et al., 2000) but the temporal stability of pleasant picture corrugator modulation remains unreported. The effect size across study visits of startle modulation to unpleasant pictures is large and has been observed to be stable (Larson et al., 2005; Lee et al., 2009), but for pleasant pictures is smaller and may be dependent on picture content (Manber et al., 2000). The effect size of corrugator modulation to unpleasant pictures is large, but may not be stable across study visits (Lee et al., 2009). The internal consistency of unpleasant and pleasant picture modulation has not been previously reported for either startle or corrugator.[2] It is also noteworthy that many of the studies reviewed here used nonstandard

1. Nelson, Hajcak, and Shankman (2015) found good internal consistency (Cronbach's alpha > .9) for startle response in each condition (no-shock, predictable shock, unpredictable shock) of the NPU task. Good internal consistency for the startle response condition scores is necessary but not sufficient to provide good internal consistency for startle potentiation (the difference between predictable or unpredictable shock conditions minus no-shock condition), which is typically the focal dependent variable of interest in the NPU task.

2. Hawk and Cook (2000) reported poor split-half reliability (r = .10) for the contrast of unpleasant minus pleasant pictures but do not evaluate modulation scores relative to neutral pictures.

implementations of the affective picture viewing task (e.g., no pleasant pictures, addition of emotion regulation instructions, contrast of same vs. different pictures). As such, conclusions about the psychometric properties of the traditional affective picture viewing task should be drawn cautiously from these studies.

Numerous studies have documented good temporal stability and internal consistency for general startle reactivity (i.e., overall startle response independent of task manipulations or stimuli; Bradley, Lang, & Cuthbert, 1993; Larson et al., 2000; Nelson, Hajcak, & Shankman, 2015; Schwarzkopf, McCoy, Smith, & Boutros, 1993). However, these evaluations have typically been conducted within the context of experimental tasks with strong manipulations and/or stimuli (e.g., affective picture viewing task, variants of the shock threat tasks) by using participant's aggregate response or responding in a putative neutral condition (e.g., intertrial interval, neutral pictures). Therefore, the psychometric properties of general startle reactivity in the resting state task, absent manipulations and experimental stimuli, remain unknown. Given recent renewed interest in general startle reactivity for both methodological and theoretical reasons (Bradford, Kaye, & Curtin, 2014; Bradford, Starr, Shackman, & Curtin, 2015; Poli & Angrilli, 2015; Vaidyanathan, Patrick, & Cuthbert, 2009), we explicitly examine this measure in the resting state task.

This brief review highlights that clear gaps and inconsistencies remain with respect to the psychometric properties of startle and corrugator response in all three of these psychophysiological tasks. Furthermore, across these studies, the quantification approach for startle (raw vs. standard scores) and corrugator response (time vs. frequency domain) has varied. Systematic evaluation and comparison of psychometric properties across common quantification approaches has been rare (but see Hawk & Cook, 2000; Larson et al., 2005; Lee et al., 2009). It is also less common to evaluate and compare both startle response and corrugator in the same task (but see Bradley et al., 1993; Lee et al., 2009; Manber et al., 2000), and we are not aware of any studies that have evaluated all tasks in the same participants.[3] Perhaps most troubling, most of these studies have relatively small sample sizes (*N*s = 20–60; but see Lee et al., 2009; Nelson et al., 2015)—an issue that has recently garnered heightened scrutiny and concern regarding the implications that small sample sizes have for the confidence and robustness of psychological and biomedical sciences (Button et al., 2013; Open Science Collaboration, 2015; Simmons, Nelson, & Simonsohn, 2011). If these psychophysiological tasks are to make an important contribution to the NIMH strategic plan (Insel et al., 2010) through the RDoC framework (Cuthbert, 2014), there is now a greater urgency to convincingly demonstrate their robust psychometric properties.

Current Study

We designed the current study to provide a comprehensive evaluation of three key psychometric properties of the three psychophysiological tasks we have described above. A large sample of participants completed the NPU task, affective picture viewing task, and resting state task at two study visits separated by approximately 1 week. We measured startle and corrugator responses in these tasks as indicators of affective processes.[4] For each task, we examine both measures with two commonly used quantification methods for each measure. Thus, we evaluated the following three psychometric properties of these three tasks, two measures, and two quantification approaches:

1. Effect size and stability: We examine the strength and stability of each focal task manipulation (e.g., unpredictable shock vs. no-shock in NPU task, unpleasant vs. neutral pictures in affective picture viewing task) by quantifying its effect size and testing for an effect of study visit (Visit 1 vs. Visit 2).
2. Internal consistency: We examine split-half reliability using Spearman-Brown–corrected Pearson correlations between odd and even trials to quantify the internal consistency within subjects.
3. Temporal stability:[5] We examine temporal stability using Pearson correlations between Study Visit 1 and Study Visit 2 to quantify the stability of individual differences in responses over 1 week.

We selected these psychometric properties because of their implications for the robustness, reproducibility, and reliability of psychological research using these tasks. Information about the effect size and stability of these task manipulations is important to avoid pitfalls (ceiling/floor effects) and to interpret change over time/sessions. Internal consistency is important to the extent that the primary variable of interest is conceptualized as representing a unidimensional construct measured homogenously by multiple trials/items. Temporal stability is an essential prerequisite for any studies aiming to make inferences about traitlike characteristics, repeat task administration within subjects under different conditions (e.g., placebo-controlled crossover designs), or detect change over time (e.g., due to intervention).

Method

Participants

We recruited 128 participants (64 female) from an introductory psychology course at the University of Wisconsin–Madison and from the greater community.[6] Participants were 18 to 61 years old (mean age = 23 years, *SD* = 7.7 years). The racial composition of the sample was 61% White, 25% Asian, 8% Black, and 6% other race. Five percent of the sample was Hispanic/Latino. We excluded participants who reported any of the following: (a) uncorrected auditory or visual problems, (b) colorblindness, (c) pregnancy, (d) current or past month use of any psychiatric medication, (e) medical or psychiatric condition that would contraindicate exposure to electric shock, or (f) severe and persistent mental illness. We compensated participants for the time they spent in the laboratory ($10/hour or course extra-credit

3. Two studies have compared startle modulation by IAPS pictures versus threat of shock. However, neither of these studies used standard versions of the NPU or affective picture viewing tasks (Bradley, Moulder, & Lang, 2005; Lissek et al., 2007).

4. To our knowledge, no previous research has established construct validity for corrugator activity in the resting state task. Therefore, we limit our analysis to raw general startle reactivity in this task.

5. We use the term *temporal stability* instead of test-retest reliability because this language more accurately describes the data analysis technique. We simply evaluate the extent to which individuals display similar responses across study visits, which neither confirms nor negates test-retest reliability. If the assumptions of parallel tests are not met, then the temporal stability of measures reflects the lower bound of test-retest reliability (DeVellis, 2012).

6. To achieve our target sample size of 128, we recruited 173 participants. Nineteen participants did not meet eligibility criteria; 17 participants withdrew from the study or did not return to second study visit. We discontinued five participants from the study. We excluded four due to errors with data acquisition.

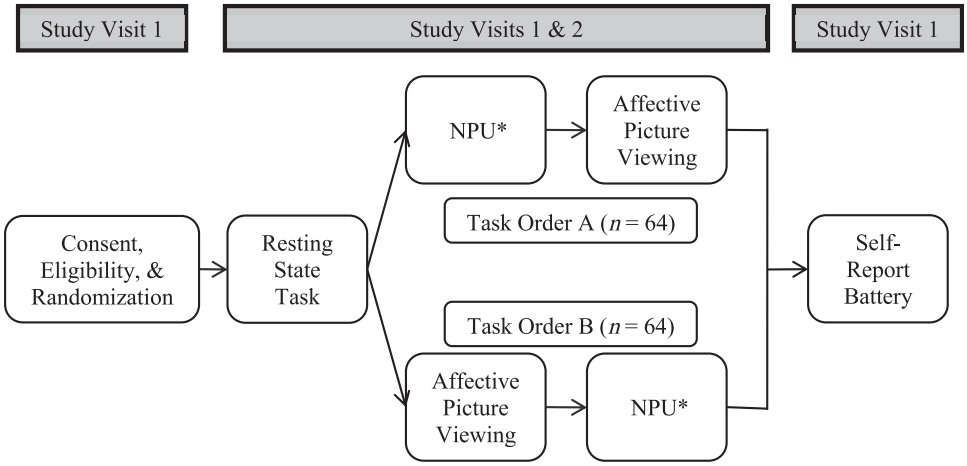

Figure 1. Study visit flowchart. Shaded boxes indicate procedures completed at each study visit. We randomly assigned eligible participants to groups based on task order (Group A: NPU first vs. Group B: NPU second) stratified by sex. Participants completed the NPU task, affective picture viewing task, and resting state task at both study visits. All participants completed the resting state task prior to the other tasks. Participants completed the same task order at both study visits. *Shock sensitivity assessment was completed at the first study visit only in order to minimize the number of shocks participants received.

points/hour) plus a bonus for completing both study visits ($20 or course extra-credit points).

General Procedures

All procedures were approved by our Institutional Review Board. All participants completed two study visits separated by approximately 1 week at approximately the same time of day. At the first study visit, we explained the study purpose and procedures, and participants provided written informed consent. We then administered surveys to evaluate inclusion/exclusion criteria. The following procedures were identical at both study visits except where noted (see Figure 1 for study procedure timeline).

We prepared participants for physiology measurements by cleaning the skin (washed with facial soap, alcohol swab, exfoliating gel) and adhering EMG sensors. Participants were comfortably seated in a dimly lit room approximately 45 inches in front of a 20-inch CRT computer monitor. Participants then completed the resting state task. Next, participants completed the NPU task and affective picture viewing tasks with a brief break between tasks. Half the participants completed the NPU task first and half completed the affective picture viewing task first. Task order was counterbalanced across participants stratified by sex and held constant at both study visits.

At the first study visit only, participants completed a battery of self-report questionnaires on an iPad (Apple Inc.) using Qualtrics software (Provo, UT) to assess demographic information, trait affect, and broadband personality traits (see online supporting information). At the second study visit only, participants repeated the resting state task at the end of the study visit. We administered these questionnaires and second resting state task for aims not relevant to the current psychometric evaluation of the tasks. Participants were debriefed, paid/compensated, and thanked for their participation.

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. We planned our sample size ($N = 128$) to be comparable to recent studies in our lab using the NPU and related tasks that are powered to detect medium between-subjects effect sizes. Sample size was a multiple of 16 to be evenly balanced across all between-subjects factors (2 Sex × 2 Task Order × 4 Trial Structures).

Resting State Task

We coded our experimental tasks in MATLAB using the Psychophysics Toolbox extensions (Kleiner et al., 2007). We measured participants' resting state startle responses prior to initiating the first experimental task (NPU or affective picture viewing) at both study visits to assess their general startle reactivity. Participants viewed a white fixation cross in the center of the black screen while nine acoustic startle probes were presented, separated by 13–20 s. No other images were displayed on the screen, and no shocks were delivered during this task. The task took approximately 2.5 min. General startle reactivity was calculated as the mean raw startle response during the resting state task.

No-Shock, Predictable Shock, Unpredictable Shock (NPU) Task

Shock sensitivity assessment. To control for individual differences in shock sensitivity, we measured participants' subjective tolerance using standard procedures from our laboratory (Bradford, Curtin, & Piper, 2015; Bradford, Magruder, Korhumel, & Curtin, 2014; Hogle et al., 2010). Participants rated a series of 200-ms electric shocks of increasing intensity (7 mA maximum) administered to the distal phalanges of the second and fourth finger of the right hand. We used participants' subjective maximum tolerated shock from this procedure during the NPU task at both study visits to minimize individual differences in subjective shock tolerance.

NPU task. During the NPU task, participants viewed a series of colored square "cues" displayed in the center of a computer screen with a black background. We presented cues in a blocked design with three conditions: no-shock (N), predictable shock (P), and unpredictable shock (U). Each shock condition was presented twice and separated by no-shock conditions. Condition order was counterbalanced both within and between subjects (i.e., two condition orders: PNUNUNP, UNPNPNU), and participants completed the same order at both study visits. All blocks included six cues presented for 5 s separated by a variable intertrial interval (ITI; mean 17 s, range 14–20 s). A white fixation cross remained in the center of the monitor during the cues and ITI. We administered a 200-ms

electric shock 200 ms prior to cue offset during every cue in the predictable shock conditions, so that the appearance of the cue "predicted" that the shock would occur in several seconds. We administered electric shock at pseudorandom times during both cues and ITIs in the unpredictable shock conditions, so that the occurrence of the shock was unpredictable by the participant. Shocks occurred 2 or 4.8 s postcue onset and 4, 8, or 12 s postcue offset in the unpredictable condition. Twelve electric shocks were administered in each predictable and unpredictable shock condition. No electric shock occurred during the no-shock condition. Each block lasted approximately 150 s, and the entire NPU task lasted approximately 20 min.

We took several steps to ensure participants clearly understood the differences between task conditions. First, we verbally instructed participants of the cue-shock contingencies and answered questions to confirm their understanding before starting the task. Second, text appeared at the top of the screen (i.e., "no shocks," "shock at end of red square," "shock at any time") for 9 s prior to the start of each block and remained throughout the entire block. Third, we disconnected the shock wire (~1 foot from fingers) prior to each no-shock condition and reconnected the wire prior to each shock condition. This provided participants with an additional signal to confirm that our shock instructions were truthful. Finally, we verified participants' task understanding and engagement by recording their verbal response to the question, "Can you be shocked in the next 5 seconds?" periodically throughout the NPU task. We told participants to answer this question (i.e., yes or no) whenever a question mark appeared on screen in place of the fixation cross. Question marks appeared on screen four times in each shock condition and six times in the no-shock condition for 3 s beginning .5 s postcue onset or 4 or 8 s postcue offset. We excluded from data analyses participants who did not answer at least 10 out of 14 questions correctly ($N = 6$).

Startle probes occurred at 4.5 s postpicture onset on a random subset of eight cues and 13, 14, or 15 s postcue offset during four ITIs in both shock conditions (no-shock condition: twelve cues and six ITIs). Startle probes occurred a minimum of 12.5 s after another startle-eliciting event (e.g., shock or startle probe). Serial position of startle probes across the three conditions for both cues and ITIs was counterbalanced within subjects to account for habituation. We used two different orders of the serial position of startle probe, counterbalanced between subjects.

Posttask subjective measures. After the NPU task, participants retrospectively reported their fear/anxiety during each condition. Means and standard deviations for these reports are provided in the supporting information Table S1.

Affective Picture Viewing Task

Affective picture viewing task. Participants viewed 36 color photographic images of unpleasant, pleasant, or neutral valence. Pictures were selected from the IAPS (Lang, Bradley, & Cuthbert, 2008). Pictures (1,024 × 768 pixels) were presented intermixed by valence for 6 s in the center of the computer monitor separated by a variable ITI (mean = 17 s, range = 14–20 s). Serial position of valence condition was counterbalanced within subjects. We used two different trial orders of the serial position of valence condition, counterbalanced between subjects. However, we presented each participant with a unique order of specific pictures within a valence condition, selected at random without replacement for each valence condition. A white fixation cross remained in the center of the monitor during the ITI. Each block lasted approximately 150 s, and the entire task lasted approximately 17 min.

We used two picture sets of 36 different pictures (12 unpleasant, 12 pleasant, 12 neutral).[7] We selected unpleasant and pleasant pictures to be high on arousal and comparably extreme on valence ratings. We selected neutral pictures to be rated low arousal and at midpoint of valence. Erotic images were overrepresented (5 of 12 per picture set) in the pleasant condition to increase startle modulation and temporal stability (Manber et al., 2000). We selected pictures to minimize differences in normative valence and arousal ratings between men and women. We used two different picture sets because 4-week temporal stability of emotion-modulated startle may be superior when participants view different rather than the same pictures (Larson et al., 2000). All participants saw both picture sets, one set at each study visit, with picture set order counterbalanced across participants. We matched Picture Set A and B on valence and arousal ratings within each condition based on normative ratings as well as picture content (e.g., people, mutilation, erotica, animals, scenery). Means and standard deviations for the normative ratings for the two picture sets are provided in supporting information Table S2.

Startle probes occurred at 3, 4, or 5 s postpicture onset on a random subset of eight pictures and 3 or 10 s postpicture offset during four ITIs in every condition. Startle probes were separated by a minimum of at least 13 s. Serial position of startle probes across the three valence conditions for both cues and ITIs were counterbalanced within subjects to account for habituation. We used two different orders of the serial position of startle probe, counterbalanced between subjects.

Posttask subjective measures. After the affective picture viewing task, participants viewed the same 36 pictures again on an iPad and rated the subjective valence and arousal of each picture. Ratings were made on a 9-point scale using the Self-Assessment Manikin (Lang et al., 2008). Means and standard deviations for participants' ratings of valence and arousal and their picture viewing times for the two picture sets are provided in supporting information Table S3.

Startle Response Measurement and Data Reduction

We recorded eyeblink EMG activity to the startle probes from two 4-mm Ag-AgCl sensors placed according to published guidelines beneath the right eye over the orbicularis oculi muscle (Blumenthal et al., 2005). An 8-mm common ground sensor was placed in the center of the forehead and a 4-mm reference sensor was placed 1 cm to the left. We filled sensors with conductive gel (ECI Electro-Gel; Electro-Cap International, Eaton, OH). We sampled EMG activity at 2500 Hz with an online band-pass filter (1–500 Hz) using NeuroScan bioamplifiers and Scan 4.5 acquisition software (Compumedics, Charlotte, NC). We reduced data offline in MATLAB using EEGLAB (Delorme & Makeig, 2004) and PhysBox plugins (Curtin, 2011).

In each task, we measured the eyeblink startle response to binaurally presented acoustic startle probes (50 ms, 102 dB white

7. IAPS numbers for Picture Set A: unpleasant: 3000, 3080, 3102, 3170, 6260, 6313, 6415, 9183, 9295, 9302, 9325, 9921; pleasant: 1710, 4641, 4650, 4680, 4690, 4695, 4698, 5700, 5833, 7270, 8030, 8502; neutral: 2200, 2230, 2381, 2440, 2480, 5510, 5740, 7006, 7010, 7020, 7035, 9070. Picture Set B: unpleasant: 3053, 3071, 3120, 3130, 6230, 6350, 9140, 9301, 9322, 9340, 9410, 9570; pleasant: 2150, 4599, 4608, 4660, 4668, 4672, 4687, 5600, 5836, 7330, 8190, 8501; neutral: 2190, 2210, 2570, 2850, 2870, 2890, 5531, 7000, 7004, 7050, 7090, 7950.

noise with near instantaneous rise time). We presented three startle probes at the start of each task to allow for stabilization of the startle response (Blumenthal et al., 2005). We did not analyze these initial three probes. Offline processing included a high-pass filter (fourth-order 28 Hz Butterworth filter, zero phase shift), creating epochs from 50-ms preprobe to 250-ms postprobe onset, and signal rectification and smoothing (second-order 30 Hz Butterworth low-pass filter, zero phase shift). We rejected trials with values greater than ± 20 μV in the 50-ms preprobe to 10-ms postprobe window as artifact (i.e., unstable baseline). We rejected trials with mean amplitude less than -10 μV in the 100–250 ms postprobe window as artifact (i.e., movement artifact and baseline overcorrection). Using the above algorithm criteria, we automatically rejected 0.3–1.1% of all trials in each task as artifact. Next, authors JTK and DEB independently visually inspected figures of epochs of all processed startle data (all trials vs. accepted trials vs. algorithm rejected trials). We collectively reviewed any figures with individual trials that we identified as atypical that were not detected by our automated rejection algorithm. We determined by consensus whether these trials were artifact if they had excessive deflection in the baseline or postprobe windows used above for automatic artifact detection. We manually rejected these individual artifact trials (0.5–1.2% of all trials in each task).[8]

We quantified the startle response as the peak amplitude 20–100 ms postprobe onset relative to a 50-ms preprobe baseline. We excluded seven participants with general startle reactivity during the resting state task at either study visit of < 5 μV (nonresponders).

We calculated startle potentiation/modulation scores with two different commonly used approaches: (1) raw scores, and (2) standardized (t score) scores (Bradford, Starr et al., 2015). For raw score startle response, we calculated the mean startle response during cues or pictures for each condition in the NPU (no-shock, predictable shock, unpredictable shock) and affective picture viewing (neutral, pleasant, unpleasant) tasks. For standardized score startle responses, we calculated the within-subject t score of the raw startle response prior to calculating condition means as above.[9] For the NPU task, we calculated startle potentiation during cues separately for unpredictable and predictable blocks as the difference between response to probes during the shock and no-shock blocks. For the affective picture viewing task, we calculated startle modulation as the difference between responses to probes during the unpleasant or pleasant pictures versus neutral pictures.

Corrugator Response Measurement and Data Reduction

We recorded facial frowning EMG activity to the pictures and shock cues from two 4-mm Ag-AgCl sensors placed according to published guidelines above the right eyebrow over the corrugator muscle (Fridlund & Cacioppo, 1986). All online data acquisition hardware/software parameters were identical to those reported above for startle response.

We measured the corrugator response to picture onset in the affective picture viewing task and cue onset in the NPU task. We quantified the corrugator response with two different commonly used approaches: (1) raw scores in the time domain, and (2) power spectral density between 28–200 Hz in the frequency domain. Offline processing in the time domain included a high-pass filter (fourth-order 28 Hz Butterworth filter, zero phase shift), signal rectification, and smoothing (fourth-order 2 Hz Butterworth low-pass filter, single pass). For both time and frequency domain, we created epochs from 1,000 ms precue or prepicture onset to 3,000 ms postpicture onset or 4,500 ms postcue onset. Offline processing in the frequency domain included calculating mean power spectral density between 28–200 Hz using Welch's method on 1-s Hamming-windowed chunks with 50% overlap, separately for the pre- and postcue/picture onset windows for each trial (Welch, 1967). For both time and frequency domain, we baseline-corrected trials by subtracting mean activity 1,000 ms precue/picture period from the entire cue/picture period. For corrugator analysis, we excluded trials when a startle probe occurred < 2 s prestimulus onset (0–4 trials per task) or trials when participants received a shock < 4.5 s postcue onset ($N = 2$ trials in NPU task unpredictable condition). We rejected trials with deflections in the time domain greater than $\pm$ 30 μV in rectified/smoothed signal (time domain processing) or $\pm$ 350 μV in raw signal (frequency domain processing) across the entire epoch window as artifact. Using the above algorithm criteria, we automatically rejected 1.6–3% of all trials as artifact across tasks in both domains. We excluded participants with $> 25\%$ of trials rejected in the NPU task (time domain $N = 3$, frequency domain $N = 4$) or affective picture viewing task (time domain $N = 2$, frequency domain $N = 3$) at either study visit. We excluded four participants with > 10 μV^2 of 60 Hz noise in the NPU task. We created average waveforms for corrugator in the time domain separately for each condition.

In both time and frequency domain, we calculated participants' responses during the picture presentation (1,000–3,000 ms) in the affective picture viewing task or the cue presentation (1,000–4,500 ms) in the NPU task. We quantified corrugator in the time domain as the maximum 500-ms mean amplitude (using a 500-ms moving window) in the participant's average waveform during the cue or picture presentation window. We quantified corrugator in the frequency domain as the mean power spectral density in the 28–200 Hz band during the entire cue or picture presentation window.

For each quantification approach, we calculated corrugator potentiation/modulation scores in the NPU task and affective picture viewing task in an identical manner to startle potentiation/modulation as described above.

Results

Data analysis was accomplished using R (R Development Core Team, 2015) with the lmSupport (Curtin, 2015) package. We report our analyses below separated by task (NPU, affective picture viewing, resting state), psychometric property (effect size and stability, internal consistency, and temporal stability), and measure (startle, corrugator). We analyze two quantification methods for startle (raw scores vs. standardized scores) and corrugator (raw scores in

8. We report the mean (SD) number of artifact-free trials that were included in each task condition in supporting information Table S8, S9 for startle and corrugator measures, respectively. We additionally report the total number of trials that could have been included in each condition average as a point of reference.

9. We calculated standardized scores using within-subject t score transformations separately for the NPU and affective picture viewing tasks. T scores were calculated separately for each study visit. Within each task and study visit, we used trial-level raw startle responses (i) to calculate participant's (j) raw startle response mean (M_j) and standard deviation (SD_j) across their trials in the task (excluding the three habituation probes). T scores were calculated as $T\text{Startle}_{ij} = ((\text{RawStartle}_{ij} - M_j)/SD_j) * 10 + 50$.

Table 1. *Effect Size and Stability for Startle Potentiation/Modulation Across Study Visits by Task and Quantification Method*

	Quantification: Raw scores			Quantification: Standardized scores		
Task: NPU	Visit 1	Visit 2	Mean	Visit 1	Visit 2	Mean
Predictable potentiation						
Magnitude	36.1*	36.9*	36.5*	9.5*	10.2*	9.8*
95% CI	[29.8, 42.4]	[30.3, 43.6]	[30.6, 42.4]	[8.4, 10.5]	[8.9, 11.4]	[8.8, 10.8]
Partial eta-squared	.544*	.529*	.580*	.740*	.700*	.765*
Unpredictable potentiation[a]						
Magnitude	26.5*	22.9*	24.7*	7.5*	6.5*	7.0*
95% CI	[21.5, 31.4]	[18.8, 27.0]	[20.5, 28.8]	[6.6, 8.5]	[5.6, 7.4]	[6.2, 7.8]
Partial eta-squared	.511*	.531*	.562*	.683*	.623*	.718*
	Quantification: Raw scores			Quantification: Standardized scores		
Task: Affective picture viewing	Visit 1	Visit 2	Mean	Visit 1	Visit 2	Mean
Pleasant modulation[b]						
Magnitude	−4.2*	−1.6	−2.9*	−1.5*	−0.1	−0.8*
95% CI	[−5.8, −2.5]	[−4.0, 0.7]	[−4.3, −1.5]	[−2.3, −0.7]	[−1.0, 0.7]	[−1.5, −0.2]
Partial eta-squared	.181*	.017	.135*	.106*	.001	.058*
Unpleasant modulation[b]						
Magnitude	6.3*	8.1*	7.2*	3.1*	4.9*	4.00*
95% CI	[4.4, 8.2]	[6.3, 9.9]	[5.7, 8.7]	[2.3, 4.0]	[4.0, 5.8]	[3.3, 4.7]
Partial eta-squared	.279*	.417*	.435*	.294*	.499*	.492*
	Quantification: Raw scores					
Task: Resting state	Visit 1	Visit 2	Mean			
General startle reactivity[a]						
Magnitude	87.3*	72.5*	79.9*			
95% CI	[75.7, 98.8]	[61.5, 83.5]	[68.9, 90.8]			
Partial eta-squared	.657*	.592*	.640*			

Note. Table cells contain effect sizes for startle potentiation (vs. no-shock) or modulation (vs. neutral picture) in magnitude (i.e., point estimate of effect from general linear model analyses in microvolts or *t*-score units depending on quantification method) and partial eta-squared for Study Visit 1, Study Visit 2, and the mean across visits for the three tasks and two quantification methods. We also report 95% confidence intervals for raw magnitude in brackets.
[a]Significant ($p < .05$) study visit effect for raw score quantification method.
[b]Significant ($p < .05$) study visit effect for standardized score quantification method.
*Significant (nonzero) effect size ($p < .05$).

the time domain vs. power spectral density in the frequency domain).[10]

We evaluated effect size and stability of scores (Objective 1) from each task, measure, and quantification method in separate general linear models (GLMs) with study visit (Visit 1 vs. Visit 2) as a within-subject factor.[11] We report both partial eta-squared (η_p^2) and raw GLM parameter estimates (*b*) to document effect sizes. We characterize effect sizes as small ($.1 \geq \eta_p^2 > .06$), moderate ($.06 \leq \eta_p^2 < .14$), or large ($\eta_p^2 \geq .14$) following established rules of thumb (Cohen, 1988). We evaluated split-half reliability of scores (Objective 2) at Study Visit 1 for each task, measure, and quantification method with Spearman-Brown–corrected correlations (r_{sb}). We evaluated the temporal stability of scores (Objective 3) for each task, measure, and quantification method with Pearson correlations (*r*) between scores from Study Visit 1 and 2. We characterize split-half reliability and temporal stability estimates as poor ($r < .5$), adequate ($.5 \leq r < .8$), or good ($r \geq .8$) based on synthesis of commonly reported thresholds for these indices (e.g., Clark & Watson, 1995; Schmitt, 1996). We provide brief summaries of all analyses in the text. Additional detail for effect size and stability analyses is provided in Table 1 and 2 for startle and corrugator measures, respectively. Additional detail for internal consistency and temporal stability is provided in Table 3 and 4 for startle and corrugator measures, respectively.[12]

10. We conducted case analyses to identify participants who were model outliers for each analysis (e.g., Bonferroni-corrected studentized residuals, $p < .05$). We removed participants listwise from all analyses within a given task/measure/quantification method to facilitate comparisons across the three psychometric properties. Based on this criterion, we dropped between zero and eight participants from each analysis. After artifact rejection and outlier removal, the final sample sizes by task, measure, and quantification method were as follows. NPU task: raw startle potentiation ($N = 110$), standardized startle potentiation ($N = 115$), time domain corrugator potentiation ($N = 105$), frequency domain corrugator potentiation ($N = 106$). Affective picture viewing task: raw startle modulation ($N = 114$), standardized startle modulation ($N = 121$), time domain corrugator modulation ($N = 120$), frequency domain corrugator modulation ($N = 116$). Resting state task: general startle reactivity ($N = 118$).

11. We included mean general startle reactivity across both study visits (mean centered) as a covariate in the analyses of raw score startle potentiation/modulation measures to control for individual differences in raw startle response as recommended (Bradford, Kaye, & Curtin, 2014; Bradford, Magruder, Korhumel, & Curtin, 2014; Schmitz & Grillon, 2012). This covariate was unnecessary for standardized scores as the standardization procedure itself is designed to reduce individual differences. Similarly, no such procedure has been proposed for corrugator response.

12. In the supporting information, we present a full description of the psychometric properties separately by each task condition (as opposed to by potentiation/modulation scores). We present the effect size and stability (Table S4, S5) and internal consistency and temporal stability (Table S6, S7) for startle and corrugator measures. While we recognize that there may be some utility to examining the task condition scores, we believe that understanding the psychometric properties of the difference scores

Table 2. *Effect Size and Stability for Corrugator Potentiation/Modulation Across Study Visits by Task and Quantification Method*

	Quantification: Raw scores in time domain			Quantification: Power in frequency domain		
Task: NPU	Visit 1	Visit 2	Mean	Visit 1	Visit 2	Mean
Predictable potentiation						
Magnitude	0.15*	0.18*	0.16*	0.015	0.020	0.017*
95% CI	[0.02, 0.28]	[0.06, 0.29]	[0.06, 0.27]	[−0.002, 0.031]	[−0.001, 0.040]	[0.002, 0.032]
Partial eta-squared	.045*	.085*	.081*	.028	.034	.045*
Unpredictable potentiation						
Magnitude	0.17*	0.18*	0.17*	0.024*	0.020*	0.022*
95% CI	[0.06, 0.27]	[0.08, 0.28]	[0.09, 0.25]	[0.007, 0.040]	[0.002, 0.038]	[0.010, 0.034]
Partial eta-squared	.093*	.115*	.154*	.074*	.044*	.109*
	Quantification: Raw scores in time domain			Quantification: Power in frequency domain		
Task: Affective picture viewing	Visit 1	Visit 2	Mean	Visit 1	Visit 2	Mean
Pleasant modulation						
Magnitude	−.01	0.04	0.02	−0.006	0.006	0.000
95% CI	[−0.13, 0.11]	[−0.10, 0.19]	[−0.88, 0.12]	[−0.027, 0.015]	[−0.029, 0.041]	[−0.023, 0.023]
Partial eta-squared	.0002	.003	.001	.003	.001	.000
Unpleasant modulation						
Magnitude	0.72*	0.83*	0.78*	0.101*	0.135*	0.118*
95% CI	[0.54, 0.91]	[0.60, 1.05]	[0.60, 0.96]	[0.062, 0.139]	[0.094, 0.177]	[0.083, 0.153]
Partial eta-squared	.340*	.309*	.379*	.189*	.266*	.278*

Note. Table cells contain effect sizes for corrugator potentiation (vs. no-shock) or modulation (vs. neutral picture) in magnitude (i.e., point estimate of effect from general linear model analyses in microvolts or power spectral density units depending on quantification method) and partial eta-squared for Study Visit 1, Study Visit 2, and the mean across visits for the two tasks and two quantification methods. We also report 95% confidence intervals for raw magnitude in brackets.
*Significant (nonzero) effect size ($p < .05$).

NPU Task

For the NPU task, we analyzed the psychometric properties of predictable and unpredictable potentiation (vs. no-shock) for both the startle and corrugator response with identical formula as described earlier in Method.

Objective 1: Effect size and stability

Startle potentiation. The predictable startle potentiation effect size was large and significant across visits for both raw scores ($\eta_p^2 = .580$, $b = 36.5\ \mu V$, $t(108) = 12.22$, $p < .001$) and standardized scores ($\eta_p^2 = .765$, $b = 9.8$, $t(114) = 19.25$, $p < .001$). The predictable startle potentiation effect size was stable across study visits (i.e., no significant effect of study visit) for both raw scores ($\eta_p^2 = .001$, $b = 0.9\ \mu V$, $t(108) = 0.33$, $p = .745$) and standardized scores ($\eta_p^2 = .014$, $b = 0.7$, $t(114) = 1.27$, $p = .208$).

The unpredictable startle potentiation effect size was large and significant for both raw scores ($\eta_p^2 = .562$, $b = 24.7\ \mu V$, $t(108) = 11.78$, $p < .001$) and standardized scores ($\eta_p^2 = .718$, $b = 7.0$, $t(114) = 17.05$, $p < .001$). The unpredictable startle potentiation effect size was stable across study visits for raw scores ($\eta_p^2 = .033$, $b = -3.6\ \mu V$, $t(108) = 1.93$, $p = .056$). However, the standardized startle potentiation effect size was significantly smaller in Visit 2 than Visit 1 ($\eta_p^2 = .040$, $b = -1.1$, $t(114) = 2.18$, $p = .031$).

Corrugator potentiation. The predictable corrugator potentiation effect size was moderate and significant for raw scores in the time domain ($\eta_p^2 = .081$, $b = 0.16\ \mu V$, $t(104) = 3.02$, $p < .003$) and small but significant in the frequency domain ($\eta_p^2 = .045$, $b = 0.017\ \mu V^2$, $t(105) = 2.22$, $p = .028$). The predictable corrugator potentiation effect size was stable across study visits in both the time domain ($\eta_p^2 = .002$, $b = 0.03\ \mu V$, $t(104) = 0.47$, $p = .637$) and frequency domain ($\eta_p^2 = .002$, $b = 0.005\ \mu V^2$, $t(105) = 0.47$, $p = .643$).

The unpredictable corrugator potentiation effect size was large and significant in the time domain ($\eta_p^2 = .154$, $b = 0.17\ \mu V$, $t(104) = 4.35$, $p < .001$) and moderate and significant in the frequency domain ($\eta_p^2 = .109$, $b = 0.022\ \mu V^2$, $t(105) = 3.57$, $p < .001$). The unpredictable corrugator potentiation effect size was stable across study visits in both the time domain ($\eta_p^2 = .001$, $b = 0.02\ \mu V$, $t(104) = 0.26$, $p = .793$) and frequency domain ($\eta_p^2 = .001$, $b = -0.003\ \mu V^2$, $t(105) = 0.28$, $p = .782$).

Objective 2: Internal consistency

Startle potentiation. Split-half reliability of predictable startle potentiation was good for raw scores ($r_{sb} = .81$) and adequate for standardized scores ($r_{sb} = .57$). Split-half reliability of unpredictable startle potentiation was adequate for raw scores ($r_{sb} = .64$) and standardized scores ($r_{sb} = .52$).

Corrugator potentiation. Split-half reliability of predictable corrugator potentiation was poor for scores in both the time domain ($r = .45$) and frequency domain ($r < .00$). Split-half reliability of unpredictable corrugator potentiation was also poor for scores in both the time domain ($r < .00$) and frequency domain ($r < .00$).

Objective 3: Temporal stability

Startle potentiation. Temporal stability of predictable startle potentiation was adequate for raw scores ($r = .71$) and standardized

relative to a control condition (e.g., neutral pictures or no-shock cues) represent the more theoretically meaningful and experimentally controlled scores of interest. It is important to emphasize that most studies that use these tasks examine difference scores in their statistical analyses either explicitly or implicitly as part of the within-subject contrasts of interest.

Table 3. *Internal Consistency and Temporal Stability of Startle Potentiation/Modulation by Task and Quantification Method*

Task: NPU	Quantification: Raw scores	Quantification: Standardized scores
Internal consistency		
Predictable potentiation[a]	.81 [.72, .87]*	.57 [.37, .70]*
Unpredictable potentiation	.64 [.48, .76]*	.52 [.31, .67]*
Temporal stability		
Predictable potentiation	.71 [.60, .79]*	.58 [.44, .69]*
Unpredictable potentiation[a]	.71 [.60, .79]*	.49 [.33, .62]*
Task: Affective picture viewing	**Quantification: Raw scores**	**Quantification: Standardized scores**
Internal consistency		
Pleasant modulation[a]	−.10 [−.38, .23]	.16 [−.17, .41]
Unpleasant modulation	.14 [−.20, .41]	.07 [−.25, .35]
Temporal stability		
Pleasant modulation	−.01 [−.19, .18]	.08 [−.10, .26]
Unpleasant modulation	.50 [.35, .63]*	.40 [.24, .54]*
Task: Resting state	**Quantification: Raw scores**	
Internal consistency		
General startle reactivity	.95 [.93, .97]*	
Temporal stability		
General startle reactivity	.89 [.85, .92]*	

Note. Table cells contain estimates of internal consistency (i.e., Spearman-Brown–corrected Pearson correlations between odd and even trials) and temporal stability (Pearson correlations between Study Visit 1 and 2) for startle potentiation (vs. no-shock), modulation (vs. neutral picture) or response (resting state) for the three tasks and two quantification methods. We also report 95% confidence intervals for these correlations in brackets.
[a]Significant difference ($p < .05$) in psychometric property (i.e., internal consistency or temporal stability) between raw and standardized score quantification methods.
*Significant (nonzero) correlation ($p < .05$).

scores ($r = .58$). Temporal stability of unpredictable startle potentiation was adequate for raw scores ($r = .71$) and poor for standardized scores ($r = .49$).

Corrugator potentiation. Temporal stability of predictable corrugator potentiation scores was adequate in the time domain ($r = .51$) but poor in the frequency domain ($r = .35$). Temporal stability of unpredictable corrugator potentiation scores were poor in both the time domain ($r = .27$) and frequency domain ($r = .00$).

Affective Picture Viewing Task

For the affective picture viewing task, we analyzed the psychometric properties of pleasant and unpleasant modulation (vs. neutral) for both the startle and corrugator response with identical formula as described earlier in the Method section.

Objective 1: Effect size and stability

Startle modulation. The pleasant startle modulation effect size was large and significant across visits for raw scores ($\eta_p^2 = .135$, $b = -2.9$ μV, $t(112) = 4.17$, $p < .001$) but small and significant for standardized scores ($\eta_p^2 = .058$, $b = -0.8$, $t(120) = 2.71$, $p = .008$). The pleasant startle modulation effect size was stable across study visits for raw scores ($\eta_p^2 = .024$, $b = 2.5$ μV, $t(112) = 1.67$, $p = .097$). However, the standardized pleasant startle modulation effect size was significantly smaller in Visit 2 than Visit 1 ($\eta_p^2 = .048$, $b = 1.4$, $t(120) = 2.45$, $p = .016$).

Table 4. *Internal Consistency and Temporal Stability of Corrugator Potentiation/Modulation by Task and Quantification Method*

Task: NPU	Quantification: Raw scores in time domain	Quantification: Power in frequency domain
Internal consistency		
Predictable potentiation[a]	.45 [.20, .63]*	−.25 [−.49, .09]
Unpredictable potentiation[a]	−.18 [−.45, .17]	−.64 [−.75, −.47]
Temporal stability		
Predictable potentiation	.51 [.35, .64]*	.35 [.17, .51]*
Unpredictable potentiation[a]	.27 [.09, .44]*	.00 [−.19, .19]
Task: Affective picture viewing	**Quantification: Raw scores in time domain**	**Quantification: Power in frequency domain**
Internal consistency		
Pleasant modulation[a]	.21 [−.12, .45]	−.46 [−.63, −.22]
Unpleasant modulation	.54 [.33, .68]*	.44 [.20, .62]*
Temporal stability		
Pleasant modulation	.20 [.02, .36]*	.30 [.12, .46]*
Unpleasant modulation	.56 [.42, .67]*	.54 [.39, .65]*

Note. Table cells contain estimates of internal consistency (i.e., Spearman-Brown–corrected Pearson correlations between odd and even trials) and temporal stability (Pearson correlations between Study Visit 1 and 2) for corrugator potentiation (vs. no-shock) or modulation (vs. neutral picture) for the two tasks and two quantification methods. We also report 95% confidence intervals for these correlations in brackets.
[a]Significant difference ($p < .05$) in psychometric property (i.e., internal consistency or temporal stability) between quantification methods of raw scores in time domain and power spectral density scores in the frequency domain.
*Significant (nonzero) correlation ($p < .05$).

The unpleasant startle modulation effect size was large and significant for both raw scores ($\eta_p^2 = .435$, $b = 7.2$ μV, $t(112) = 9.29$, $p < .001$) and standardized scores ($\eta_p^2 = .492$, $b = 4.0$, $t(120) = 10.78$, $p < .001$). The unpleasant startle modulation effect size was stable across study visits for raw scores ($\eta_p^2 = .026$, $b = 1.8$ μV, $t(112) = 1.74$, $p = .084$). However, standardized unpleasant startle modulation was significantly larger in Visit 2 than Visit 1 ($\eta_p^2 = .098$, $b = 1.8$, $t(120) = 3.61$, $p < .001$).

Corrugator modulation. The pleasant corrugator modulation effect size was not significant in either the time domain ($\eta_p^2 = .001$, $b = 0.017$ μV, $t(119) = 0.33$, $p = .746$) or frequency domain ($\eta_p^2 = .000$, $b = 0.000$ μV^2, $t(115) = 0.01$, $p = .992$). The pleasant corrugator modulation effect size was stable across study visits for both the time domain ($\eta_p^2 = .003$, $b = 0.054$ μV, $t(119) = 0.62$, $p = .537$) and frequency domain ($\eta_p^2 = .004$, $b = 0.012$ μV^2, $t(115) = 0.68$, $p = .499$).

The unpleasant corrugator modulation effect size was large and significant in both the time domain ($\eta_p^2 = .379$, $b = 0.777$ μV, $t(119) = 8.52$, $p < .001$) and frequency domain ($\eta_p^2 = .278$, $b = 0.118$ μV^2, $t(115) = 6.66$, $p < .001$). The unpleasant corrugator modulation effect size was stable across study visits for both the time domain ($\eta_p^2 = .009$, $b = 0.104$ μV, $t(119) = 1.06$, $p = .293$) and frequency domain ($\eta_p^2 = .027$, $b = 0.035$ μV^2, $t(115) = 1.78$, $p = .078$).

Objective 2: Internal consistency

Startle modulation. Split-half reliability was poor for pleasant startle modulation for both raw scores ($r_{sb} < .00$) and standardized scores ($r_{sb} = .16$). Split-half reliability was also poor for unpleasant startle modulation for both raw scores ($r_{sb} = .14$) and standardized scores ($r_{sb} = .07$).

Corrugator modulation. Split-half reliability was poor for pleasant corrugator modulation scores in both the time domain ($r_{sb} = .21$) and frequency domain ($r_{sb} < .00$). Split-half reliability for unpleasant corrugator modulation was adequate for scores in the time domain ($r_{sb} = .54$) and poor for scores in the frequency domain ($r_{sb} = .44$).

Objective 3: Temporal stability

Startle modulation. Temporal stability of pleasant startle modulation was poor for raw scores ($r < .00$) and standardized scores ($r = .08$). Temporal stability of unpleasant startle modulation was adequate for raw scores ($r = .50$) and poor for standardized scores ($r = .40$).

Corrugator modulation. Temporal stability of pleasant corrugator modulation was poor for scores in both the time ($r = .20$) and frequency domains ($r = .30$). Temporal stability of unpleasant corrugator modulation was adequate for scores in both the time ($r = .56$) and frequency domains ($r = .54$).

Resting State Task

For the resting state task, we only evaluate the psychometric properties of general startle reactivity (i.e., mean startle response) quantified as raw scores as described earlier in the Method section. Within-subject standardized scores have no utility (i.e., all scores would be 50).

The raw score general startle reactivity effect size was smaller in Visit 2 than Visit 1 ($\eta_p^2 = .209$, $b = -14.82$, $t(117) = 5.55$, $p < .001$). The split-half reliability of raw score general startle reactivity was good ($r = .95$). The temporal stability of raw score general startle reactivity was high ($r = .89$).

Discussion

We conducted a rigorous and systematic evaluation of three key psychometric properties of three psychophysiology tasks. These tasks are situated prominently within the RDoC negative valence system and have a wide range of potential applications from the study of individual differences to mechanistic experimental medicine trials. We measured both startle and corrugator response within these tasks, used two distinct quantification methods for each measure, and evaluated their performance in a relatively large sample that allowed more precise estimates of these psychometric properties than many previously published reports. In the sections that follow, we first consider implications of the psychometric properties of each task separately. We conclude with discussion of general issues across tasks, limitations, and associated directions for future research.

NPU Task

The startle response has been the focal dependent measure in the NPU task across multiple laboratories (Moberg & Curtin, 2009; Schmitz & Grillon, 2012; Shankman et al., 2013). This decision appears well justified with respect to its psychometric properties documented here. Predictable and unpredictable shock threat produce robust, exceptionally large potentiation of the startle reflex across both raw score (25–37 μV) and standard score (7–10 t units) quantification methods. In variance terms, the partial eta-squared effect sizes range from .56–.77, with > .14 considered a large effect by convention. Effect sizes were not statistically compared across quantification methods, but the variance estimates were modestly higher descriptively for standard scores than raw scores. Nonetheless, shock threat appears to potently potentiate this defensive reflex regardless of quantification method.

Predictable and unpredictable startle potentiation also displayed adequate to good internal consistency ($r_{sb} = 0.5$–0.8) across quantification methods. There was some evidence that raw scores may be more internally consistent than standard scores. However, the combination of robust, large effects with adequate internal consistency suggests that the NPU task, with either startle potentiation quantification approach, is well suited for studies that require a single administration of the task.

The NPU task also appears appropriate for research questions that require multiple administrations across time. At the group level, the effect sizes for both predictable and unpredictable startle potentiation were generally stable across two administrations separated by 1 week. The only exception to this observation was noted from unpredictable potentiation quantified by standard scores, where a small but significant decrease was detected for the second administration. Nonetheless, potentiation scores remain exceptionally large for both methods at Study Visit 2 (η_p^2s ranged from 0.53–0.70), which eliminates any concern about floor effects that might arise if the defensive response to the shock threats habituated across repeated administrations of the task. Of course, this study does not allow us to definitively evaluate the impact of additional (beyond two) administrations separated by more time. However, given the size and stability of the potentiation scores across two administrations, we expect that additional administrations at later visits are likely feasible.

Reliable measurement of change across administrations requires more than the simple demonstration of effect size stability in the overall sample. Within-subject temporal stability in the relative size of effects is also necessary. We observed the temporal stability of predictable and unpredictable potentiation to be appropriate for this purpose as well, with correlations between Visit 1 and 2 ranging from 0.5–0.7 across quantification methods. These results align closely with a previous report on the temporal stability of startle potentiation in the NPU task by Shankman et al. (2013). As with internal consistency, we observed some evidence that the temporal stability of startle potentiation was higher for raw than standard scores. However, all correlations were significant and at least of adequate size, which we believe supports the use of the NPU task in pre–post and other designs to assess efficacy of manipulations expected to produce within-person change.

To our knowledge, there is no previously published report using corrugator to assess reactivity to predictable and unpredictable shock in the NPU task. In this study, predictable and unpredictable shock threat produced significant corrugator potentiation of varying effect size across quantification methods (η_p^2s ranged from .05–.15). Although significant and in some conditions moderate to large in size, these effects are noticeably smaller than those reported above for startle response. The internal consistency of corrugator potentiation was generally very poor. In fact, only predictable potentiation in the time domain displayed a significant, positive, Spearman-Brown–corrected correlation between odd and even trials ($r_{sb} = .45$), which only approached levels considered adequate for internal consistency (i.e., $> .50$). Furthermore, the temporal stability within individuals was also generally poor. As with internal consistency, predictable potentiation in the time domain displayed the highest temporal stability ($r = .51$) but only at the threshold to be considered adequate ($> .50$). In aggregate, it appears that significant corrugator potentiation in the NPU task may result from only a small subset of task trials within and across participants. These critical trials may be adequate to detect a threat effect in the full sample. However, there is reasonable concern that heterogeneity in reactivity across trials within participants and across administrations of the task will yield low power both to detect effects of other focal manipulations in a single task administration and to detect change in reactivity within participants over repeated task administrations.

Conclusions. Predictable and unpredictable startle potentiation in the NPU task appears well suited for both single administration and longitudinal or other research designs with multiple administrations (e.g., pre–post designs, drug vs. placebo crossover designs). There is some evidence that startle potentiation raw scores are modestly superior to standard scores in their internal consistency and temporal stability of effect size within participants and across the entire sample. However, the effect sizes may be modestly larger for standard scores. Clearly, more systematic comparison of these two quantification methods is warranted (Bradford, Starr et al., 2015). Corrugator potentiation appears adequate to detect predictable and unpredictable threat reactivity, but concerns with internal consistency and temporal stability within and across participants combined with smaller effects than observed for startle may limit the utility of measuring corrugator in this task.

Affective Picture Viewing Task

The startle response has been a focal dependent measure in the affective picture viewing task across multiple laboratories (Donohue, Curtin, Patrick, & Lang, 2007; Lang et al., 2008; Larson et al., 2005; Vaidyanathan, Patrick, & Bernat, 2009). Mean pleasant and unpleasant startle modulation across study visits was significant for both raw and standard score quantification methods. However, these modulatory effects varied substantially in size, with effects generally larger for unpleasant (η_p^2s ranged from .47–.49) than pleasant pictures (η_p^2s ranged from .06–.14). Effect sizes did not appear to vary systematically by quantification method.

In contrast to the generally robust mean effect sizes across visits, concerns were noted for other psychometric properties of pleasant and unpleasant startle modulation. Poor internal consistency was observed for both pleasant and unpleasant modulation across quantification methods. Specifically, all Spearman-Brown–corrected correlations were small and nonsignificant, suggesting substantial heterogeneity in response across trials such that effects might be only carried by a small subset of trials/pictures. Significant study visit effects were detected for both pleasant and unpleasant modulation when quantified by standard scores, suggesting that these effect sizes were not stable across task administrations in the full sample. Study visit effects were not observed for raw scores. However, there is some evidence that pleasant modulation may not persist across administrations given that significant pleasant modulation was not observed at Study Visit 2 for either raw or standard score methods. The temporal stability of unpleasant modulation was adequate for raw ($r = .50$) but poor for standard scores ($r = .40$) with significant Pearson correlations between scores across study visits. The temporal stability was poor for pleasant modulation, with nonsignificant Pearson correlations across study visits for both methods. Given these concerns about internal consistency and temporal stability, it is not surprising that a recent rigorous study found no genetic associations to suggest that startle modulation is heritable in this task (Vaidyanathan, Malone, Miller, McGue, & Iacono, 2014). It may be difficult to detect any stable, genetic and/or environmental trait variation with startle in this task.

Corrugator is also frequently measured within the affective picture viewing task (Jackson, Malmstadt, Larson, & Davidson, 2000; Larsen et al., 2003; Manber et al., 2000). This appears reasonably justified for unpleasant corrugator modulation based on its psychometric properties. Significant, large unpleasant mean corrugator modulation across study visits was observed for both quantification methods (η_p^2s = .38 and .28 in time and frequency domains, respectively). The Spearman-Brown–corrected correlations for internal consistency of unpleasant modulation were both significant and adequate ($r = .54$) for scores in the time domain but poor ($r = .44$) for scores in the frequency domain. The robust, unpleasant modulation effect was stable across study visits in the full sample for both methods, with significant modulation observed at both study visits for both methods. The temporal stability of unpleasant modulation was also adequate with significant positive Pearson correlations between scores across study visits (rs = .56 and .54 in time and frequency domains, respectively).

In contrast, there is serious concern about the use of pleasant corrugator modulation in the affective picture viewing task. No significant pleasant modulation was observed for either individual study visits or the mean across visits for either method. Internal consistency was poor, with no significant, positive Spearman-Brown correlations observed for either method. Similarly, within-participant temporal stability was poor with no significant, positive Pearson correlations between scores across study visits observed for either method.

Conclusions. Both pleasant and unpleasant startle modulation is very heterogeneous across trials/pictures such that effects may depend on a few key pictures. As such, picture selection may be very important. Poor internal consistency for startle modulation may also limit its sensitivity to detect effects of other manipulations and the reproducibility of these other effects across studies. Unpleasant pictures appear to produce more robust modulation of both startle and corrugator that persists over study visits relative to pleasant pictures in this task. Pleasant pictures may not be useful for situations that require repeated task administration due to small/null effects for subsequent administrations and the absence of any temporal stability across measures. A number of studies have now consistently demonstrated poor psychometric performance of startle/corrugator modulation to pleasant pictures, which raises serious concerns regarding the conclusions that can be drawn about pleasant picture responding for either measure (Bradley et al., 2001; Larson et al., 2000, 2005; Manber et al., 2000). Unpleasant but not pleasant corrugator modulation has comparable psychometric properties to startle modulation. With the exception of some instability in effect sizes across study visits for standardized startle modulation, the performance of two quantification methods evaluated for each measure was generally comparable.

Resting State Task

There has been recent, renewed interest in the measurement of general startle reactivity both as an individual difference construct of interest and for potential methodological benefits associated with its use as a covariate (Bradford, Kaye, & Curtin, 2014; Bradford, Starr et al., 2015; Poli & Angrilli, 2015; Vaidyanathan, Patrick, & Bernat, 2009). Although general startle reactivity can be measured in a variety of task contexts, its measurement within the resting state task appears to yield strong psychometric properties. Specifically, raw score general startle reactivity had both high internal consistency ($r = .95$) and within-participant temporal stability ($r = .89$). The significant study visit effect indicates that it does habituate to some degree over administrations (see also Larson et al., 2005). However, robust overall responding was still observed in Study Visit 2, which eliminates any concern about floor effects preventing measurement of the response over repeated administrations of the task.

Conclusions. General startle reactivity possesses admirable internal consistency and temporal stability within subjects. It is clearly well suited for measurement in experiments that require single or repeated administration. Its high reliability provides a solid foundation for its use as a covariate in analyses of startle modulation/potentiation (Bradford, Kaye, & Curtin, 2014). Other research has recently demonstrated that it is highly heritable (Vaidyanathan et al., 2014), which combined with its traitlike temporal stability suggests that it may have use as a dispositional measure within the RDoC negative valence system. Further examination of its construct validity is clearly warranted.

General Issues, Limitations, and Future Directions

We have generally avoided explicit comparisons of the psychometric properties across the NPU and affective picture viewing tasks. Although both tasks can be used to index psychological constructs within the negative valence system, it is likely that the tasks and their associated measures and potentiation/modulation scores may tap distinct mechanisms within this domain. The NPU task putatively measures response to acute threat (fear) and potential harm (anxiety) constructs. It is less clear where to situate response to unpleasant pictures in the affective picture viewing task within this domain. Future research with our and other datasets can use multitrait/multimethod approaches to quantify convergent and discriminant validity among measures (startle, corrugator) and potentiation/modulation scores within these two tasks and relevant self-report measures. Such efforts would also align well with Patrick and colleagues' "psychoneurometric" approach to combine psychophysiological tasks, self-report measures, and clinical interviews to more powerfully index underlying latent constructs (Patrick et al., 2013). If appropriate, construct measurement with a battery of tasks/measures using varied methods may offer improved psychometric function beyond that offered by each task/measure in isolation.

Nonetheless, some differences in the psychometric properties of the NPU and affective picture viewing tasks appear so robust as to deserve brief mention. Both predictable and unpredictable threat of shock appear to increase the startle response more potently than unpleasant pictures. This suggests that actual direct physical threat (e.g., shock) is a stronger manipulation of defensive neural circuits than unpleasant pictures (Lissek et al., 2007). This raises concerns about the ability of unpleasant pictures to robustly activate defensive systems that modulate the startle response. Conversely, corrugator response is more strongly modulated by unpleasant pictures than threat of shock, perhaps due to the "social" nature of a subset of unpleasant pictures (Larsen et al., 2003). These considerations when pairing task and measure may be relevant when concerns exist about the impact of floor (or ceiling) effects and stimulus potency on task sensitivity to detect effects of other manipulations (e.g., drug administration) and/or individual differences (Bradford, Starr et al., 2015; Lissek, Pine, & Grillon, 2006).

The internal consistency of startle within the NPU was substantially higher than in the affective picture viewing task. This may not be surprising given that shock threat is essentially identical on every trial within condition but picture content is highly variable within picture valence. Indeed, picture diversity may increase the construct validity of the affective picture viewing task by sampling more broadly across stimuli that elicit affect. However, when startle modulation in the affective picture viewing task displays essentially no internal consistency (i.e., no significant Spearman-Brown–corrected correlations for pleasant or unpleasant modulation by either quantification method), concerns about picture selection and reproducibility become more fundamental than sampling the breadth of the construct. This concern might be reduced by increasing the number of trials within the affective picture viewing task, though the Spearman-Brown prediction formula suggests that the number of trials would need to be increased between 5–13 times the length we used to achieve even marginal internal consistency.[13] It may be that the internal consistency of startle modulation

13. The Spearman-Brown prophecy formula can be used to predict the required increase in test length to achieve a desired reliability: $N = (r_{desired} * (1 - r_{observed}))/(r_{observed} * (1 - r_{desired}))$. If we use the mean split-half reliability across quantification methods for pleasant and unpleasant modulation (observed $r_{sb} = .07$), the number of startle probe trials would have to be increased by thirteenfold to achieve adequate split-half reliability (desired $r_{sb} = .60$). If we use the highest split-half reliability we observed (observed $r_{sb} = .16$), the number of probe trials would have to be increased fivefold to achieve the same desired reliability.

scores is higher with valence subcategories (e.g., threat, disgust, erotica). However, our study was not designed to evaluate this as it would require many more trials within each subcategory. Future studies that present more homogeneous valence subcategories may indeed yield higher internal consistency.

The striking low internal consistency for raw startle modulation during unpleasant pictures may at first seem at odds with the adequate temporal stability we observed for this measure. However, poor internal consistency and reasonable temporal stability can arise in a task where only small subsets of trials both discriminate between people and possess good temporal stability. Low internal consistency results when the correlations across trials within individuals is generally low. However, if a few of these trials discriminate well between individuals (i.e., have high variance) and these same trials are temporally stable, the overall measure will possess adequate temporal stability even though it has low internal consistency. The IAPS may be an example of a stimulus set where the responses across pictures within a valence category do not correlate highly, but a few pictures discriminate well between individuals and are temporally stable within these same individuals. Nonetheless, this is problematic because it makes picture selection difficult and critical effects less likely to replicate if different studies use different picture sets.

More generally, we recommend that researchers regularly report the internal consistency of their measures in their manuscript even if psychometric evaluation is not the primary aim of the study (Patrick & Hajcak, 2016). Indeed, it is commonplace to report the Cronbach's alpha or comparable metric to document the internal consistency of self-report measures. There is no reason that psychophysiology tasks should not be subject to the same standards.

We did compare quantification methods within measure to some degree. Specifically, we explicitly tested differences in the correlations used to quantify internal consistency and temporal stability. When significant differences between the sizes of these correlations were noted for startle, they always favored raw over standard scores. Furthermore, all significant correlations, individually, were descriptively larger for raw than standard scores. Standard startle scores were also more likely than raw scores to display instability in the size of potentiation/modulation effects over study visits for the NPU and affective picture viewing tasks, though it is not clear if this is artifact or real effect. This study was not designed to arbitrate between these two startle quantification methods but combines with other recent research to amplify calls for more direct, systematic, rigorous comparisons between these two methods (Bradford, Starr et al., 2015). For corrugator, there were fewer significant and meaningful differences between quantification methods in the time and frequency domains. However, this may be because, with the exception of unpleasant modulation in the affective picture viewing task, corrugator potentiation/modulation performance was generally suboptimal in these tasks.

We chose to focus on the psychometric properties of the potentiation/modulation (i.e., difference) scores in the NPU and affective picture viewing tasks rather than the properties of the individual condition scores (e.g., startle response during unpleasant pictures alone, corrugator response during predictable shock alone). We believe this is appropriate because the potentiation/modulation scores index the fundamental constructs of interest (e.g., fear, reactivity to threat) in the vast majority of studies, either explicitly (i.e., when startle potentiation difference scores are calculated and used in the analysis) or implicitly by testing single degree of freedom contrasts for a within-subject condition variable (e.g., unpleasant vs. neutral picture startle response) in the analytic model. Temporal stability and internal consistency of condition scores alone are necessary but not sufficient to achieve desirable psychometric properties for potentiation/modulation scores or contrasts that represent the constructs of interest. As such, the psychometric properties of these potentiation/modulation scores must be verified directly as we have done in this paper.

We characterized the psychometric properties of these tasks primarily by reference to the sample parameter estimates from our study. Of course, all parameter estimates have sampling error associated with them. For this reason, we took care to report 95% confidence intervals around all parameter estimates in our tables. We believe this is important to better characterize the range of possible population values for these psychometric properties, but report of confidence intervals has not been routinely employed previously. We collected a sample size that is large relative to many previously published psychometric studies of these tasks, thus the confidence intervals we report are likely smaller than many previously published reports. However, they remain at times wide in our sample. Statistical tests of these parameter estimates against zero include the uncertainty reflected in these intervals. However, decisions about the utility of these tasks and measures based on their psychometric properties rest more on the actual magnitude of these parameters in the population, not simply if they are nonzero. As such, the reader should carefully consult these confidence intervals when making judgments about the range of likely population values for these psychometric properties. In some instances, they do span values that indicate favorable versus unfavorable task performance.

With respect to data acquisition, processing, and screening, we followed guidelines on best practices for each task and measure when available (Blumenthal et al., 2005; Bradford, Kaye, & Curtin, 2014; Fridlund & Cacioppo, 1986; Schmitz & Grillon, 2012). We also believe we were particularly rigorous in screening for artifact at trial and participant levels. As such, the psychometric properties reported here may more closely represent an upper bound if future users are less thorough. Regardless, we provide substantial detail in the Method and supporting information to assure our decisions are transparent and reproducible.

With respect to data reduction, we focused on commonly used potentiation/modulation scores within each task with two quantification methods for each measure to increase the generalizability of our findings. Nonetheless, others might prefer to use alternative quantification methods (e.g., standardized corrugator response in the time domain) or to calculate different contrasts altogether (e.g., unpredictable potentiation in intertrial interval rather than cue period in the NPU task, unpleasant vs. pleasant pictures in affective picture viewing task). Following Open Science recommendations (Nosek et al., 2015), we provide unrestricted access to all raw data and analysis code from this study to the research community via Open Science Framework (https://osf.io/fdjg9/). As such, others will be able to examine alternative quantification methods and contrasts with relative ease. While broad but useful guidelines exist for many aspects of data acquisition, processing, and screening, we hope that additional research with these tasks leads to definitive recommendations regarding likely impactful decisions about quantification methods. Otherwise, excessive "researcher degrees of freedom" regarding these decisions may undermine the reproducibility of results with these tasks (Simmons et al., 2011). We take this special issue in *Psychophysiology* as yet another strong indicator that our field both values and critically evaluates the rigor of our tasks and measures.

References

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology*, *42*(1), 1–15. doi: 10.1111/j.1469-8986.2005.00271.x

Bradford, D. E., Curtin, J. J., & Piper, M. E. (2015). Anticipation of smoking sufficiently dampens stress reactivity in nicotine-deprived smokers. *Journal of Abnormal Psychology*, *124*(1), 128–136. doi: 10.1037/abn0000007

Bradford, D. E., Kaye, J. T., & Curtin, J. J. (2014). Not just noise: Individual differences in general startle reactivity predict startle response to uncertain and certain threat. *Psychophysiology*, *51*(5), 407–411. doi: 10.1111/psyp.12193

Bradford, D. E., Magruder, K. P., Korhumel, R. A., & Curtin, J. J. (2014). Using the threat probability task to assess anxiety and fear during uncertain and certain threat. *Journal of Visualized Experiments*, *91*. doi: 10.3791/51905

Bradford, D. E., Starr, M. J., Shackman, A. J., & Curtin, J. J. (2015). Empirically based comparisons of the reliability and validity of common quantification approaches for eyeblink startle potentiation in humans. *Psychophysiology*, *52*(12), 1669–1681. doi: 10.1111/psyp.12545

Bradley, M. M., Codispoti, M., Cuthbert, B., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion*, *1*(3), 276–298.

Bradley, M. M., & Lang, P. J. (2007). The International Affective Picture System (IAPS) in the study of emotion and attention. In J. A. Coan & J. J. B. Allen (Eds.), *Handbook of emotion elicitation and assessment*. Oxford, UK: Oxford University Press.

Bradley, M. M., Lang, P. J., & Cuthbert, B. N. (1993). Emotion, novelty, and the startle reflex: Habituation in humans. *Behavioral Neuroscience*, *107*(6), 970–980.

Bradley, M. M., Moulder, B., & Lang, P. J. (2005). When good things go bad: The reflex physiology of defense. *Psychological Science*, *16*(6), 468–473. doi: 10.1111/j.0956-7976.2005.01558.x

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi: 10.1038/nrn3475

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. doi: 10.1037/1040-3590.7.3.309

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J: Routledge.

Curtin, J. J. (2011). *PhysBox: The psychophysiology toolbox. An open source toolbox for psychophysiological data reduction within EEGLAB*. Retrieved from http://dionysus.psych.wisc.edu/PhysBox.htm

Curtin, J. J. (2015). lmSupport: Support for linear models (Version 2.9.2). Retrieved from https://cran.r-project.org/web/packages/lmSupport/index.html

Cuthbert, B. N. (2014). Response to Lilienfeld. *Behaviour Research and Therapy*, *62*, 140–142. doi: 10.1016/j.brat.2014.08.001

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

DeVellis, R. (2012). Scale development theory and applications. In *Scale Development Theory and Applications* (3rd ed., Vol. 26). New York, NY: Sage Publications, Inc.

Donohue, K. F., Curtin, J. J., Patrick, C. J., & Lang, A. R. (2007). Intoxication level and emotional response. *Emotion*, *7*(1), 103–112.

Fridlund, A. J., & Cacioppo, J. T. (1986). Guidelines for human electromyographic research. *Psychophysiology*, *23*(5), 567–589. doi: 10.1111/j.1469-8986.1986.tb00676.x

Grillon, C., Baas, J. M., Pine, D. S., Lissek, S., Lawley, M., Ellis, V., & Levine, J. (2006). The benzodiazepine alprazolam dissociates contextual fear from cued fear in humans as assessed by fear-potentiated startle. *Biological Psychiatry*, *60*(7), 760–766.

Grillon, C., Lissek, S., Rabin, S., McDowell, D., Dvir, S., & Pine, D. S. (2008). Increased anxiety during anticipation of unpredictable but not predictable aversive stimuli as a psychophysiologic marker of panic disorder. *American Journal of Psychiatry*, *165*(7), 898–904. doi: 10.1176/appi.ajp.2007.07101581

Grillon, C., Pine, D. S., Lissek, S., Rabin, S., Bonne, O., & Vythilingam, M. (2009). Increased anxiety during anticipation of unpredictable aversive stimuli in posttraumatic stress disorder but not in generalized anxiety disorder. *Biological Psychiatry*, *66*(1), 47–53. doi: 10.1016/j.biopsych.2008.12.028

Hajcak, G., & Patrick, C. J. (2015). Situating psychophysiological science within the Research Domain Criteria (RDoC) framework. *International Journal of Psychophysiology*, *98*(2, Pt. 2), 223–226. doi: 10.1016/j.ijpsycho.2015.11.001

Hawk, L. W., & Cook, E. W. (2000). Independence of valence modulation and prepulse inhibition of startle. *Psychophysiology*, *37*(1), 5–12.

Heller, A. S., Greischar, L. L., Honor, A., Anderle, M. J., & Davidson, R. J. (2011). Simultaneous acquisition of corrugator electromyography and functional magnetic resonance imaging: A new method for objectively measuring affect and neural activity concurrently. *NeuroImage*, *58*(3), 930–934. doi: 10.1016/j.neuroimage.2011.06.057

Hogle, J. M., Kaye, J. T., & Curtin, J. J. (2010). Nicotine withdrawal increases threat-induced anxiety but not fear: Neuroadaptation in human addiction. *Biological Psychiatry*, *68*(8), 687–688. doi: 10.1016/j.biopsych.2010.06.003

Insel, T. R. (2015). The NIMH experimental medicine initiative. *World Psychiatry*, *14*(2), 151–153. doi: 10.1002/wps.20227

Insel, T. R., 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. *American Journal of Psychiatry*, *167*(7), 748–751. doi: 10.1176/appi.ajp.2010.09091379

Insel, T. R., & Gogtay, N. (2014). National Institute of Mental Health clinical trials: New opportunities, new expectations. *JAMA Psychiatry*, *71*(7), 745–746. doi: 10.1001/jamapsychiatry.2014.426

Jackson, D. C., Malmstadt, J. R., Larson, C. L., & Davidson, R. J. (2000). Suppression and enhancement of emotional responses to unpleasant pictures. *Psychophysiology*, *37*(4), 515–522. doi: 10.1111/1469-8986.3740515

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Lang, P., Bradley, M. M., & Cuthbert, B. (2008). International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual. University of Florida, Gainsville, FL.

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*(5), 776–785.

Larson, C. L., Ruffalo, D., Nietert, J. Y., & Davidson, R. J. (2000). Temporal stability of the emotion-modulated startle response. *Psychophysiology*, *37*(1), 92–101.

Larson, C. L., Ruffalo, D., Nietert, J. Y., & Davidson, R. J. (2005). Stability of emotion-modulated startle during short and long picture presentation. *Psychophysiology*, *42*(5), 604–610. doi: 10.1111/j.1469-8986.2005.00345.x

Lee, H., Shackman, A. J., Jackson, D. C., & Davidson, R. J. (2009). Test-retest reliability of voluntary emotion regulation. *Psychophysiology*, *46*(4), 874–879. doi: 10.1111/j.1469-8986.2009.00830.x

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

Lissek, S., Orme, K., McDowell, D. J., Johnson, L. L., Luckenbaugh, D. A., Baas, J. M., . . . Grillon, C. (2007). Emotion regulation and potentiated startle across affective picture and threat-of-shock paradigms. *Biological Psychology*, *76*(1-2), 124–133. doi: 10.1016/j.biopsycho. 2007.07.002

Lissek, S., Pine, D. S., & Grillon, C. (2006). The strong situation: A potential impediment to studying the psychobiology and pharmacology of anxiety disorders. *Biological Psychology*, *72*(3), 265–270.

Manber, R., Allen, J. J., Burton, K., & Kaszniak, A. W. (2000). Valence-dependent modulation of psychophysiological measures: Is there consistency across repeated testing? *Psychophysiology*, *37*(5), 683–692.

Miller, G. A., & Rockstroh, B. (2013). Endophenotypes in psychopathology research: Where do we stand? *Annual Review of Clinical Psychology*, *9*(1), 177–213. doi: 10.1146/annurev-clinpsy-050212-185540

Moberg, C. A., & Curtin, J. J. (2009). Alcohol selectively reduces anxiety but not fear: Startle response during unpredictable vs. predictable threat. *Journal of Abnormal Psychology*, *118*(2), 335–347. doi: 10.1037/a0015636

Nelson, B. D., Hajcak, G., & Shankman, S. A. (2015). Event-related potentials to acoustic startle probes during the anticipation of predictable and unpredictable threat. *Psychophysiology*, *52*(7), 887–894. doi: 10.1111/psyp.12418

NIMH. (2015). Research Domain Criteria (RDoC). Retrieved from http://www.nimh.nih.gov/research-priorities/rdoc/index.shtml

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*, *348*(6242), 1422–1425. doi: 10.1126/science.aab2374

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

Patrick, C. J., & Hajcak, G. (2016). RDoC: Translating promise into progress. *Psychophysiology*, *53*(3), 415–424. doi: 10.1111/psyp.12612

Patrick, C. J., Venables, N. C., Yancey, J. R., Hicks, B. M., Nelson, L. D., & Kramer, M. D. (2013). A construct-network approach to bridging diagnostic and physiological domains: Application to assessment of externalizing psychopathology. *Journal of Abnormal Psychology*, *122*(3), 902–916. doi: 10.1037/a0032807

Poli, E., & Angrilli, A. (2015). Greater general startle reflex is associated with greater anxiety levels: A correlational study on 111 young women. *Frontiers in Behavioral Neuroscience*, *9*, 10. doi: 10.3389/fnbeh.2015.00010

R Development Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. doi: 10.1037/1040-3590.8.4.350

Schmitz, A., & Grillon, C. (2012). Assessing fear and anxiety in humans using the threat of predictable and unpredictable aversive events (the NPU-threat test). *Nature Protocols*, *7*(3), 527–532. doi: 10.1038/nprot.2012.001

Schwarzkopf, S. B., McCoy, L., Smith, D. A., & Boutros, N. N. (1993). Test-retest reliability of prepulse inhibition of the acoustic startle response. *Biological Psychiatry*, *34*(12), 896–900.

Shankman, S. A., Nelson, B. D., Sarapas, C., Robison-Andrew, E. J., Campbell, M. L., Altman, S. E., ... Gorka, S. M. (2013). A psychophysiological investigation of threat and reward sensitivity in individuals with panic disorder and/or major depressive disorder. *Journal of Abnormal Psychology*, *122*(2), 322–338. doi: 10.1037/a0030747

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Vaidyanathan, U., Malone, S. M., Miller, M. B., McGue, M., & Iacono, W. G. (2014). Heritability and molecular genetic basis of acoustic startle eye blink and affectively modulated startle response: A genome-wide association study. *Psychophysiology*, *51*(12), 1285–1299. doi: 10.1111/psyp.12348

Vaidyanathan, U., Patrick, C. J., & Bernat, E. M. (2009). Startle reflex potentiation during aversive picture viewing as an indicator of trait fear. *Psychophysiology*, *46*(1), 75–85. doi: 10.1111/j.1469-8986.2008.00751.x

Vaidyanathan, U., Patrick, C. J., & Cuthbert, B. N. (2009). Linking dimensional models of internalizing psychopathology to neurobiological systems: Affect-modulated startle as an indicator of fear and distress disorders and affiliated traits. *Psychological Bulletin*, *135*(6), 909–942. doi: 10.1037/a0017222

Welch, P. D. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, *15*(2), 70–73. doi: 10.1109/TAU.1967.1161901

(Received November 30, 2015; Accepted March 22, 2016)

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix S1: Self-report battery and supplemental results.

Table S1: Post-NPU task subjective self-report measures.

Table S2: IAPS picture normative mean male and female valence and arousal ratings.

Table S3: IAPS picture study sample mean male and female valence and arousal ratings.

Table S4: Effect size and stability for startle response across study visits.

Table S5: Effect size and stability for corrugator response across study visits.

Table S6: Internal consistency and temporal stability of startle response.

Table S7: Internal consistency and temporal stability of corrugator response.

Table S8: Number of valid trials analyzed for startle response across study visits.

Table S9: Number of valid trials analyzed for corrugator response across study visits.