International Journal of Psychophysiology 114 (2017) 24–30

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

How many blinks are necessary for a reliable startle response? A test using the NPU-threat task☆

Lynne Lieberman [a], Elizabeth S. Stevens [a], Carter J. Funkhouser [a], Anna Weinberg [b], Casey Sarapas [a], Ashley A. Huggins [c], Stewart A. Shankman [a,*]

[a] University of Illinois at Chicago, Department of Psychology, Chicago, IL 60657, United States
[b] McGill University, Department of Psychology, Montreal, QB, Canada
[c] University of Wisconsin, Department of Psychology, Milwaukee, WI 53211, United States

ARTICLE INFO

Article history:
Received 19 July 2016
Received in revised form 17 January 2017
Accepted 31 January 2017
Available online 2 February 2017

Keywords:
Anxiety-potentiated startle
Emotion-modulated startle
Eyeblink startle reflex
Fear-potentiated startle
Reliability

ABSTRACT

Emotion-modulated startle is a frequently used method in affective science. Although there is a growing literature on the reliability of this measure, it is presently unclear how many startle responses are necessary to obtain a reliable signal. The present study therefore evaluated the reliability of startle responding as a function of number of startle responses (NoS) during a widely used threat-of-shock paradigm, the NPU-threat task, in a clinical ($N = 205$) and non-clinical ($N = 92$) sample. In the clinical sample, internal consistency was also examined independently for healthy controls vs. those with panic disorder and/or major depression and retest reliability was assessed as a function of NoS. Although results varied somewhat by diagnosis and for retest reliability, the overall pattern of results suggested that six startle responses per condition were necessary to obtain acceptable reliability in clinical and non-clinical samples during this threat-of-shock paradigm in the present study.

1. Introduction

Establishing the reliability of a measure is an essential first step towards establishing its validity (Cronbach, 1947; Cronbach and Meehl, 1955). Although this fact is well accepted in the development of self-report and interview measures, the psychometric properties of psychophysiological indices of psychological constructs has received less attention until recently (Hajcak and Patrick, 2015; Tomarken, 1995). This is particularly important given the increasingly prominent role of psychophysiological measures within psychology (and affective science more specifically; Schwartz et al., 2016; Shankman and Gorka, 2015). The present study therefore seeks to contribute to this burgeoning literature by examining the reliability of a widely used psychophysiological index of emotion – electromyography of the eyeblink startle reflex (EMG startle).

The startle reflex is particularly conducive to translational research on emotion because it is present across species and its magnitude is modulated by an organism's emotional state. More specifically, the magnitude of the startle reflex is potentiated or blunted relative to baseline when an organism is in an aversive (e.g., fear) or appetitive (e.g., excitement) emotional state, respectively (Grillon and Ameli, 2001; Vrana et al., 1988). Startle is also commonly used to examine emotional processing abnormalities that may contribute to the development and maintenance of psychopathology. For example, heightened aversive responding to particular threatening stimuli/situations has been implicated in the pathogenesis of several internalizing disorders (e.g., panic disorder and interoceptive cues; posttraumatic stress disorder and trauma-related cues; social anxiety disorder and social evaluation; Craske et al., 2009). However, *unpredictable* threatening stimuli are particularly aversive for anxious individuals. Panic disorder (PD), posttraumatic stress disorder, and social anxiety disorder have all been associated with heightened startle potentiation during the anticipation of unpredictable threat (Cornwell et al., 2006; Grillon et al., 2009; Shankman et al., 2013). Thus, aberrant emotion-modulated startle, particularly during the anticipation of unpredictable threat, may represent a transdiagnostic marker for several internalizing disorders.

The literature on the psychometric properties of emotion-modulated startle has also grown in recent years. Investigations of the retest reliability of emotion-modulated startle elicited during an affective picture-viewing task have yielded mixed results, with some investigations finding strong retest reliability (Bradley et al., 1995; Larson et al., 2000) and others finding weak retest reliability (Kaye et al., 2016; Manber et al., 2000). Only two studies to date have examined retest reliability of emotion-modulated startle during the No threat-Predictable threat-Unpredictable threat-task (NPU; Schmitz and Grillon, 2012), a startle paradigm that is widely used to differentiate startle potentiation

☆ Funding: This study was supported by grants from the National Institute of Mental Health (S.S., grant numbers R01 MH098093 and R21 MH080689).

* Corresponding author at: University of Illinois at Chicago, 1007 W. Harrison St. (M/C 285), Chicago, IL 60657, United States.
E-mail address: stewarts@uic.edu (S.A. Shankman).

http://dx.doi.org/10.1016/j.ijpsycho.2017.01.012
0167-8760/

to predictable threat (i.e., fear-potentiated startle) and unpredictable threat (i.e., anxiety potentiated startle). Both studies reported retest correlations above 0.69 for anxiety-potentiated startle and fear-potentiated startle (Kaye et al., 2016; Shankman et al., 2013). Kaye et al. (2016) reported acceptable internal consistency (i.e., Cronbach's alphas > 0.70 [Nunnally, 1978]) for anxiety-potentiated startle and fear-potentiated startle during the NPU-threat task.

Despite growing focus in the field of psychology on exploring the reliability of emotion-modulated startle, there are several major gaps in the extant literature on the psychometric properties of this psychophysiological measure. For example, it is presently unknown how many startle responses are necessary to obtain a reliable index of startle potentiation scores during emotion-modulated startle paradigms. It is also presently unknown whether the number of startle responses (NoS) necessary for reliable condition averages (which are used to calculate startle potentiation scores) and potentiation scores differs for those with internalizing psychopathology relative to those without. This is a particularly important question to address given the abovementioned association between internalizing psychopathology and aberrant emotion-modulated startle.

Condition averages and potentiation scores calculated from a sufficient NoS should demonstrate acceptable internal consistency and strong retest reliability. Determining the *minimum* number of startle responses (NoS) necessary for reliable condition averages and potentiation scores would be highly beneficial for the design of future experimental protocols (at least with the NPU startle paradigm), which should be as brief as possible to reduce participant burden and the potential impact of startle habituation on task effects (Blumenthal et al., 2005). An empirically determined minimum NoS could also help experimenters determine when a participant has too few usable startle responses to be included in data analyses. This is critical given that certain trials may be excluded for some participants due to artifacts (e.g., excessive participant movement just before or after the presentation of a startle probe) and non-responses (i.e., failure to exhibit a discernable startle response) and some participants may withdraw from the study prior to study completion.

Several studies have examined the reliability of event-related potentials as a function of number of trials (e.g., Foti et al., 2013; Moran et al., 2013; Meyer et al., 2013). However, only one study to our knowledge has examined this question with respect to EMG startle data. Our laboratory recently investigated the NoS necessary for adequate internal consistency (i.e., degree of interrelatedness or stability; Tavakol and Dennick, 2011) of average startle magnitude during each condition of the NPU-threat task (i.e., condition averages) in a non-clinical sample. Startle magnitude exhibited excellent internal consistency (Cronbach's alpha > 0.80) for all NPU conditions with as few as three responses (Nelson et al., 2015). The present study will replicate our previous investigation by examining the internal consistency of condition averages during NPU as a function of NoS across *two* additional samples, one clinical and one non-clinical. We will also extend our previous investigation by examining; (a) the internal consistency of potentiation scores (i.e., fear-potentiated startle and anxiety potentiated startle) as a function of NoS; and (b) whether the NoS necessary for adequate consistency of condition averages and potentiation scores differs for those with an anxiety and/or depressive disorder. Lastly, we will conduct exploratory analyses to assess the NoS necessary for significant retest reliability of condition averages and potentiation scores in a subset of participants.

2. Methods

2.1. Participants

Data from the present study was collected as part of two investigations on emotional and cognitive processes. Details of the two studies are provided elsewhere (see Sarapas et al., 2017; Shankman et al., 2013). In brief, Study 1 ($n = 92$) was a non-clinical sample of undergraduates. Study 2 ($n = 205$) was a clinical sample recruited from the community to be in one of four groups: (1) no history of Axis I psychopathology (i.e., healthy controls; $n = 82$), (2) current major depressive disorder (MDD) and no lifetime history of any anxiety disorder (i.e., MDD-only group; $n = 37$), (3) current PD and no lifetime history of MDD (i.e., PD-only group; $n = 28$), (4) current PD and MDD (i.e., comorbid PD and MDD group; $n = 58$). Diagnoses were made via the Structured Clinical Interview for DSM-IV (SCID; First et al., 1996).

Exclusion criteria for both studies were a history of head trauma, left-handedness, and English fluency. Participants in Study 2 were additionally required to have no lifetime history of a psychotic disorder, bipolar disorder, or dementia. Participant demographics can be found in Table 1, along with clinical characteristics, such as self-reported anxiety and depressive symptomology.

2.2. Procedure and NPU-threat task

The full procedure for Studies 1 and 2 has been reported elsewhere (Sarapas et al., 2017; Shankman et al., 2013). In brief, after informed consent all participants completed the NPU threat-task. For Study 2, 34 participants returned to the laboratory 5–17 ($M = 9.46$, $SD = 3.71$) days after their initial visit to complete NPU a second time. Of these 34 individuals, 7 had MDD-only, 5 had PD-only, 10 had comorbid PD and MDD, and 12 were healthy controls. All procedures were approved by the local Institutional Review Board.

The NPU-threat task was designed to assess responses to predictable and unpredictable threats (Schmitz and Grillon, 2012). In brief, prior to the task, shock electrodes were placed on participants' left wrist and a shock work-up procedure was completed to identify the level of shock intensity each participant described as “highly annoying but not painful” (between 1 and 5 mA). Participants also completed a 2-min startle habituation task prior to the task to reduce early, exaggerated startle potentiation.

The NPU-threat task included three within-subjects conditions - no shock (N), predictable shock (P), and unpredictable shock (U). Text at the bottom of the computer monitor informed participants of the current threat condition and each condition lasted for 90 s. In Study 1, a 6-s countdown was displayed five times within each condition, and in Study 2, an 8-s geometric cue (blue circle for N, red square for P, and green star for U) was presented four times within each condition. Interstimulus intervals ranged from 7 to 17 s during which only the text describing the condition was on the screen (i.e., ISI conditions).

During N, no shocks were delivered. During P, Study 1 participants only received a shock when the countdown reached 1 and Study 2 participants only received a shock when the cue (red square) was on the screen (i.e., the shock was predicted by the countdown or cue in Studies 1 and 2, respectively). In the U condition, shocks were administered at any time (i.e., during the cue countdown [hereafter: cue] or ISI). Study 1 participants received 20 shocks (10 each during P and U) and 48 startle probes (16 each during N, P, and U). Study 2 participants received 12 shocks (6 during P and 6 during U) and 72 startle probes (24 each during N, P, and U). Study 2's NPU was divided into two recording blocks, separated by a rest period.

Table 1
Sample demographics and clinical characteristics.

Characteristic	Clinical sample	Non-clinical sample
Age	32.93 (12.31)	19.02 (1.38)
Gender (% female)	64.40	76.1
Ethnicity (% Caucasian)	46.30	35.9
IDAS-dysphoria	22.26 (10.61)	21.74 (81.90)
IDAS-panic	11.93 (5.34)	11.78 (4.00)
IUS-12	28.22 (10.09)	27.74 (8.67)

Note. IDAS = Inventory for Depression and Anxiety Symptoms (Watson et al., 2007); IUS-12 = Intolerance of Uncertainty scale (Carleton et al., 2007).

Stimuli (i.e., shocks, white noise) were administered using PSYLAB (Contact Precision Instruments, London, UK) hardware and software. Psychophysiological data were acquired using Neuroscan 4.4 (Compumedics, Charlotte, NC). Acoustic startle probes were 40 ms, 103-dB bursts of white noise presented binaurally through headphones. Electric shocks were 400 ms. Consistent with published guidelines (Blumenthal Blumenthal et al., 2005), EMG startle was recorded from two 4-mm Ag/AgCl electrodes placed over the orbicularis oculi muscle below the right eye and the ground electrode was at the frontal pole (AFZ). Data were collected using a bandpass filter of DC to 200 Hz at a sampling rate of 1000 Hz.

Startle blinks were scored according to published guidelines (Blumenthal et al., 2005). Data processing included applying a 28 Hz high-pass filter, rectifying, and then smoothing using a 40 Hz low-pass filter. Blink response was defined as the peak amplitude of EMG activity within the 20–150 ms period following startle probe onset relative to baseline. The baseline period was defined as the average baseline EMG level for the 50 ms preceding the startle probe onset. Each peak was identified by software but examined by hand to ensure acceptability. Blinks were scored as nonresponses if EMG activity during the 20–150 ms poststimulus time frame did not produce a blink peak that was visually differentiated from baseline activity. Blinks that were scored as nonresponses were included as zeros. Blinks were scored as missing if the baseline period was contaminated with noise, movement artifact, or if a spontaneous or voluntary blink began before minimal onset latency and thus interfered with the startle probe-elicited blink response.

2.3. Data analysis plan

Reliability was examined separately for Studies 1 and 2. Reliability was also examined separately for startle amplitude (non-responses scored as missing values) and magnitude (non-responses scored as zeros). Cronbach's alpha was used to index internal consistency (Santos, 1999). We first examined Cronbach's alpha as a function of the NoS entered into the averages for each condition (N_{Cue}, P_{Cue}, U_{Cue}, N_{ISI}, P_{ISI}, and U_{ISI}) with a maximum of 8 (Study 1) and 12 (Study 2) probes per condition. Condition averages were derived from raw microvolt values. For each NoS (NoS = 2; NoS = 3, etc.), startle probes were selected in the order that they occurred in (i.e., sequentially).[1] Given that, as mentioned above, some startle responses were scored as missing during EMG data processing, it is important to note that the available sample size of participants for all reliability analysis decreased as the NoS increased. The median number of probes that elicited missing responses was 2 (out of 48) for Studies 1 and 4 (out of 72) for Study 2 and the median number of non-responses was 1 in each sample (see Tables 2 and 3).[2] Also of note is that no case analyses were conducted, and no model outliers were removed. That is, all participants who completed the NPU-threat task in each study were included in the analyses.

Internal consistency analyses were conducted separately for each diagnostic group for Study 2 (i.e., healthy controls, PD-only, MDD-only, and comorbid MDD/PD). Cronbach's alpha was defined as 'acceptable' when equal to or >0.70 (Nunnally, 1978). Split-half reliability analyses were conducted to examine the internal consistency of potentiation scores as a function of NoS. To do so, averages of odd-numbered trials and even-numbered trials were first separately calculated as a function of NoS (e.g., the average of startle responses one and three; the average of startle responses two and four, etc.). Spearman-Brown corrected Coefficients were then calculated to assess the relation between odd and even trials (see Kappenman et al., 2014 and Kaye et al., 2016 for a similar approach). Consistent with the literature, Spearman-Brown Coefficients were interpreted as acceptable if >0.50 (Kaye et al., 2016).

For Study 2, retest reliabilities were tested as a function of NoS for: (1) average startle in each of the six NPU conditions, (2) startle potentiation to the unpredictable threat (average U_{Cue} minus average N_{Cue} and average U_{ISI} minus average N_{ISI}), and (3) startle potentiation to the predictable threat (average P_{Cue} minus average N_{Cue}). Pearson's *r* was also used to assess retest reliability.

Table 2
Sample size at each NoS for the non-clinical sample's Cronbach's alpha analyses of magnitude condition averages.

NoS	N_{ISI}	N_{Cue}	P_{ISI}	P_{Cue}	U_{ISI}	U_{Cue}
2	80	81	72	80	80	83
3	77	73	66	72	74	79
4	74	71	60	68	69	76
5	72	66	56	67	66	73
6	68	64	53	64	63	67
7	66	61	50	61	63	63
8	64	58	50	60	58	61

Note. NoS = Number of startle responses; N = No shock condition; P = Predictable shock condition; U = Unpredictable shock condition; ISI = Inter-stimulus interval.

3. Results

3.1. Internal consistency in the non-clinical sample (Study 1)

At only two responses (NoS = 2), Cronbach's alphas for average startle magnitude ranged from 0.70 to 0.83 for all conditions (see Fig. 1A). For average startle amplitude with two responses, Cronbach's alphas were comparable, ranging from 0.79 to 0.86 for all conditions except P_{Cue} (0.68). Cronbach's alpha for amplitude during P_{Cue} reached an acceptable level of 0.75 at three responses. For magnitude and amplitude potentiation scores, Spearman-Brown Coefficients reached an acceptable level across all conditions at just two responses total (range of *rs* = 0.73–0.86 and *rs* = 0.71–0.86, respectively, $p < 0.05$ [see Fig. 1B]).

3.2. Internal consistency in the clinical sample (Study 2)

Across all four groups, at two responses, Cronbach's alphas for startle magnitude and amplitude ranged from 0.85 to 0.90 across all six conditions (see Fig. 1C). Similarly, for magnitude and amplitude potentiation scores, split-half correlations reached an acceptable level across all conditions at just two responses (range of *rs* = 0.85–0.86 for magnitude and amplitude, $p < 0.05$ [see Fig. 1D]).

Table 3
Sample size at each NoS for the clinical sample's Cronbach's alpha analyses of magnitude condition averages.

NoS	N_{ISI}	N_{Cue}	P_{ISI}	P_{Cue}	U_{ISI}	U_{Cue}
2	178	172	187	193	187	187
3	169	159	183	188	175	166
4	152	139	173	180	168	156
5	135	128	156	166	159	151
6	130	119	147	155	145	143
7	122	109	142	147	139	137
8	116	105	135	141	133	134
9	110	102	129	136	129	126
10	102	99	122	130	126	123
11	100	94	118	119	115	118
12	97	88	111	109	108	112

Note. NoS = Number of startle responses; N = No shock condition; P = Predictable shock condition; U = Unpredictable shock condition; ISI = Inter-stimulus interval.

[1] The pattern of results was comparable when internal consistency analyses were conducted by adding startle responses to reliability estimates in a random order. For this method, at each NoS (NoS = 2; NoS = 3, etc), startle probes were randomly selected from all possible non-missing startle probes. For example, for NoS = 3, if a participant in study two had all 12 non-missing startle probes for a condition, 3 of the 12 were randomly selected for the analyses.

[2] The median is more appropriate than the mean in this context as 'number of missings' and 'number of nonresponses' were highly skewed (i.e., the vast majority of probes elicited startle responses).

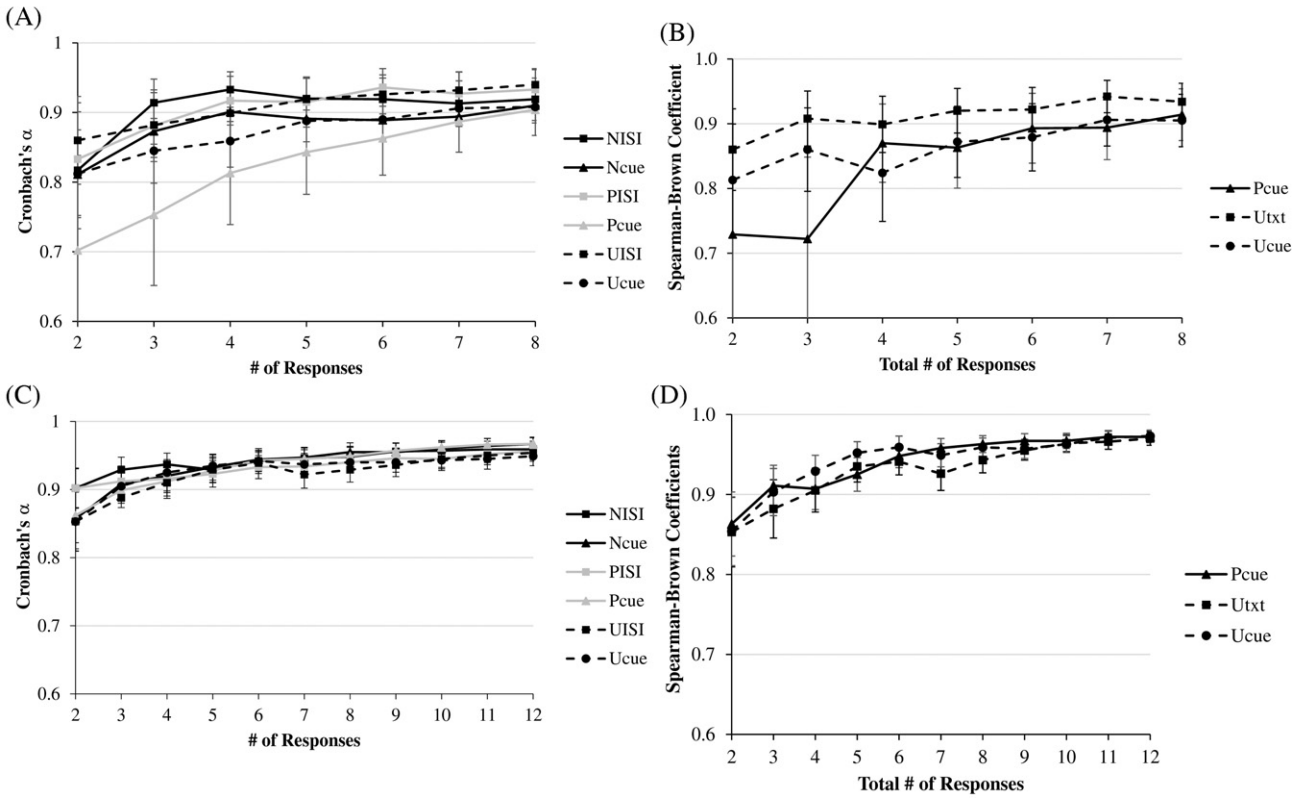

Fig. 1. Internal consistency, as indexed by Cronbach's alpha, of startle magnitude as a function of number of responses during each condition of the NPU-threat task in the (A) non-clinical, and (C) clinical sample (across all diagnostic groups). Split-level correlations as a function of responses for potentiation scores in the (B) non-clinical, and (D) clinical sample (across all diagnostic groups). Error bars represent a 95% confident interval.

The number of responses necessary to reach acceptable Cronbach's alpha levels across all conditions was comparable across diagnostic groups. In healthy controls alphas across all conditions ranged from 0.86 to 0.90 for magnitude (see Fig. 2A) and 0.85 to 0.90 for amplitude at NoS = 2. In the MDD-only group, alphas across all conditions ranged from 0.78 to 0.92 for magnitude (see Fig. 2B) and 0.77 to 0.91 for amplitude at NoS = 2. For startle amplitude in the PD-only group, alphas ranged from 0.80 to 0.90 across all conditions except N_{ISI} at NoS = 2. Likewise, for startle magnitude in the PD-only group, alphas ranged for from 0.82 to 0.90 across all conditions except N_{ISI} at NoS = 2. Alpha for magnitude and amplitude during N_{ISI} reached an acceptable level of 0.81 at NoS = 3 (see Fig. 2C). Lastly, in the comorbid MDD/PD group, alphas across all conditions ranged from 0.82 to 0.94 for magnitude (see Fig. 2D) and 0.83 to 0.93 for amplitude at NoS = 2.

Given that alpha values for magnitude and amplitude were acceptable for all conditions across all diagnostic groups at NoS = 3, exploratory follow-up analyses were conducted to examine whether Cronbach's alpha values significantly differed between those with a diagnosis of PD and/or MDD relative to healthy controls. To compare internal consistency estimates at this NoS between individuals with and without a diagnosis, Cronbach's alpha values at NoS = 3 were calculated for individuals with *any* diagnosis (i.e., collapsing across individuals with PD-only, MDD-only, or comorbid PD/MDD). We then conducted a series of pairwise comparisons using a dependent-alpha calculator developed by Abd-El-Fattah and Hassan (2011) to statistically compare Cronbach's alpha at NoS = 3 for individuals with any diagnosis, relative to healthy controls for the key threat conditions of the NPU-threat task: P_{Cue}, U_{Cue}, and U_{ISI}. These comparisons revealed no significant differences between Cronbach's alpha values at NoS = 3 for individuals with a diagnosis, relative to those without.

3.3. Retest reliability (Study 2)

For all conditions except N_{Cue} and P_{ISI}, there was a significant positive retest correlation for startle magnitude across the two visits with as few as NoS = 2 (range of *rs* = 0.38–0.71, *ps* < 0.05, see Fig. 3A). Retest correlations for startle magnitude reached significance for N_{Cue} and P_{ISI} at NoS = 3 (*rs* = 0.28 and 0.31, respectively, $p < 0.05$). Similarly, for all conditions except N_{Cue} and P_{ISI}, there was a positive retest correlation for startle amplitude with as few as NoS = 2 (range of *rs* across conditions at three responses = 0.44–0.78, *ps* < 0.05). Retest correlations for N_{Cue} startle amplitude reached significance at NoS = 5 ($r = 0.39$, $p < 0.05$), and P_{ISI} at NoS = 3 ($r = 0.35$, $p < 0.05$).

Startle potentiation to unpredictable threat during visit one positively predicted startle potentiation during visit two with as few as two startle responses for magnitude (*rs* for U_{Cue} and U_{ISI} at two responses = 0.61 and 0.49, respectively, $p < 0.05$) and amplitude (*rs* for U_{Cue} and U_{ISI} at two responses = 0.59 and 0.56, respectively, $p < 0.05$). Retest reliability for P_{Cue} reached significance at NoS = 6 for amplitude ($r = 0.38$, $p < 0.05$) and magnitude ($r = 0.36$, $p < 0.05$ [Fig. 3B]).

4. Discussion

EMG of emotion-modulated startle is a commonly used index of emotional processing and startle potentiation to threat has been used as a measure of heightened negative emotional responding to threatening stimuli/situations in various anxiety disorders (Cornwell et al., 2006; Grillon et al., 2009). Given the potential for emotion-modulated startle to serve as a transdiagnostic marker of multiple internalizing conditions, there is a growing literature on the psychometric properties of this psychophysiological measure. This is the first study, however, to

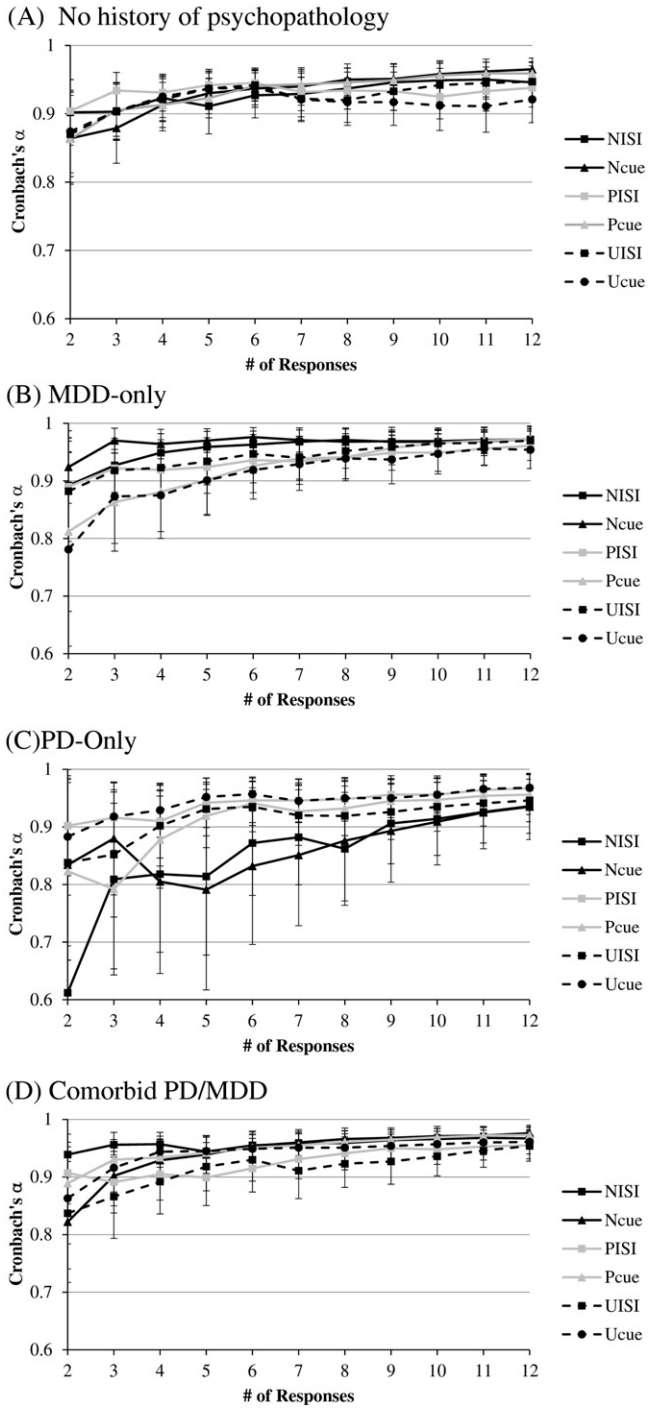

Fig. 2. (A) No history of psychopathology (B) MDD-only (C) PD-Only (D) Comorbid PD/MDD Note. Internal consistency, as indexed by Cronbach's alpha, of startle magnitude as a function of number of responses during each condition of the NPU-threat task in the clinical sample among individuals with (A) no history of psychopathology, (B) MDD-only, (C) PD-only and (D) comorbid PD/MDD. Error bars represent a 95% confident interval.

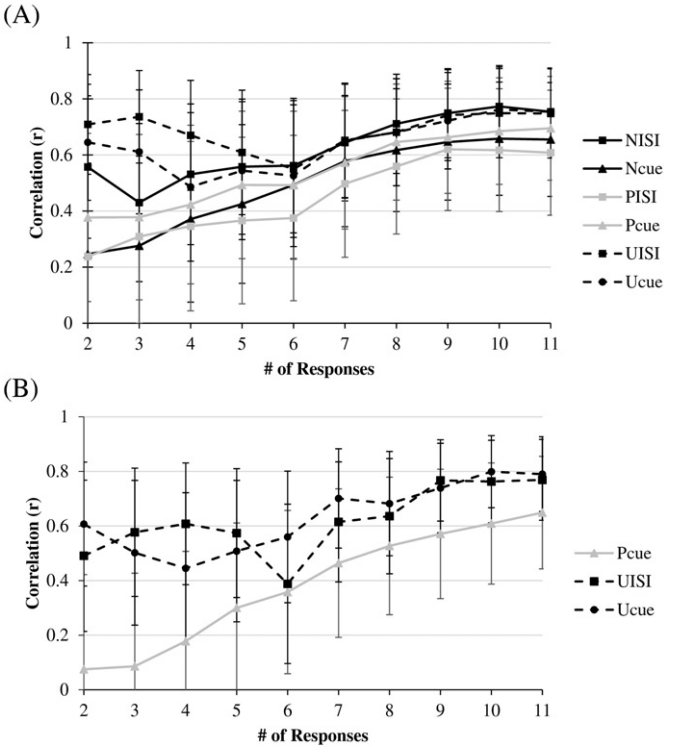

Fig. 3. Retest reliability in the clinical sample, as indexed by Pearson's r, of average startle magnitude during (A) each condition of the NPU-threat task, as well as (B) startle magnitude potentiation to predictable and unpredictable threats (P_{Cue}–N_{Cue}, U_{Cue}–N_{cue}, U_{ISI}–N_{ISI}). Error bars represent a 95% confident interval.

examine the reliability of EMG startle as a function of number of startle responses during each condition of the NPU-threat task, a widely used threat of shock paradigm, in two samples – one clinical and one non-clinical. In the clinical sample, we also explored retest reliability in a smaller subset of subjects as a function of number of startle responses for: (1) NPU condition averages, (2) anxiety-potentiated startle to unpredictable threat (U_{ISI}/U_{Cue}), and (3) fear-potentiated startle to predictable threat (P_{Cue}).

In the non-clinical sample, two responses were necessary for magnitude and three responses for amplitude condition averages to reach acceptable internal consistency (alpha > 0.70) across all conditions. This pattern of results is similar to our laboratory's previous finding that as few as two responses were necessary for magnitude to reach acceptable internal consistency across all NPU conditions in a non-clinical sample (Nelson et al., 2015). In the clinical sample, just two startle responses were necessary for condition averages (for magnitude and amplitude) to reach acceptable internal consistency across all conditions. Importantly, the internal consistency results for condition averages were similar across MDD-only, PD-only, comorbid MDD/PD, healthy controls, suggesting that internalizing psychopathology did not negatively impact reliability. Internal consistency of startle potentiation to threat, a commonly used index of negative emotional responding, was comparable to that of condition averages. More specifically, split-half correlations for magnitude and amplitude startle potentiation scores reached an acceptable level across all threat conditions in the non-clinical and clinical samples at just two responses total.

Of note is that the NoS necessary for significant retest reliability of average startle and potentiation scores differed between task conditions. All condition averages exhibited significant retest reliability at just two responses except for P_{ISI} and N_{Cue}. For P_{ISI} and N_{Cue} to exhibit significant retest reliability for amplitude and magnitude, five responses were necessary. As safety conditions in a threatening task, P_{ISI} and N_{Cue} may elicit greater variability and inconsistency in startle responding within a given task administration than do clearly threatening conditions (Lissek et al., 2006). Retest reliability reached significance at just two responses for amplitude and magnitude potentiation to U_{Cue} and U_{ISI}. However, retest reliability did not reach significance for P_{Cue} until NoS = 6, suggesting that startle potentiation to predictable threat may be somewhat more variable than to unpredictable threat.

It is noteworthy that reactivity to unpredictable threat may be more reliable than reactivity to predictable threat, as the literature on the relation between startle potentiation to *predictable* threat and anxiety

psychopathology is less consistent (e.g., Shankman et al., 2013; Grillon et al., 2008) than the literature on the relation between startle potentiation to *unpredictable* threat and anxiety psychopathology (e.g., Gorka et al., 2017; Lieberman et al., 2017; Shankman et al., 2013). That is, mixed findings on the relation between anxiety psychopathology and reactivity to predictable threat may be in part due to the poorer reliability of startle potentiation during the anticipation of predictable threat. It is also noteworthy that a higher NoS was necessary for significant retest reliability of P_{Cue} relative to the NoS necessary for acceptable internal consistency of P_{Cue}. This suggests that researchers may need to obtain a greater number of startle responses for temporal stability of startle potentiation to predictable threat, whereas fewer responses may be necessary for internal consistency of startle condition averages during P_{Cue}. Relatedly, researchers may place a greater emphasis on the results from retest analyses when designing a study that aims to obtain a temporally stable index of startle. Temporally stable indices of startle may be particularly relevant in clinical research, which may use startle responding as a predictor of risk for psychopathology or response to treatment for psychopathology.

In interpreting retest reliability results, however, it is important to consider several factors. First, this was an exploratory analyses conducted in a smaller sample ($n = 34$). Second, although retest correlations reached statistical significance for the majority of conditions at just two responses, the coefficients were moderate at this NoS. Retest correlations increased in magnitude as the NoS increased. This pattern of results suggests that the retest reliability of startle condition averages and potentiation scores is improved by a greater NoS.

In sum, investigators may only need *six* startle responses in non-clinical and clinical samples to obtain reliable and stable indices of average startle amplitude or magnitude in each condition of NPU, as well as of anxiety-potentiated and fear-potentiated startle during NPU. It is worth noting that potentiation scores (rather than startle during the individual conditions) are often the metric of interest in the NPU-threat task and other emotion-modulated startle paradigms. Given this, it is encouraging for psychophysiological researchers that so few startle responses were necessary for potentiation scores and the condition averages that are used to calculate those potentiation scores. As mentioned above, compared to self-report and interview measures of psychological variables, the psychometrics of psychophysiological tasks are often ignored, but this pattern has begun to change. For example, there have been recent investigations on how best to quantify startle potentiation or change within a paradigm (Bradford et al., 2015). Moreover, Kaye et al. (2016) investigated the internal consistency of startle condition averages and potentiation scores. Results from the present study are consistent with those reported by Kaye et al., 2016, such that startle during the NPU-threat task exhibited acceptable internal consistency and temporal stability. Furthermore, NoS analyses reported here suggest that the significant retest reliability reported by Shankman et al. (2013) in this same clinical sample and, could have been obtained with half as many startle responses. There have also been recent investigations to determine the number of events necessary to obtain reliable ERP averages (Foti et al., 2013; Moran et al., 2013; Meyer et al., 2013). Results from ERP investigations of this nature yielded results that are similar to that of the present study, such that a minimum of seven and eight responses have been suggested to obtain a reliable index of the late positive potential and error-related negativity, respectively. This exploratory study therefore adds to this growing methodological literature, and provides an empirically determined guideline to consider when developing a task to assess for emotion-modulated startle (or at least with the NPU paradigm).

Given that startle probes are naturally aversive and participant startle responses tend to habituate over the course of a task (Blumenthal et al., 2005; Campbell et al., 2014), it is important that researchers design their startle tasks to be as brief as possible to decrease participant burden and increase the quality of the psychophysiological data collected. Although data from the present study suggests that a minimum of six may be sufficient to obtain reliable and stable indices of startle during NPU, it is important to note that several responses were excluded from analyses after data collection due to artifacts or non-responses. For the non-clinical sample in the present study, a median of two responses was scored as missing and one as non-response (out of 48 responses across six conditions). For the clinical sample, a median of four responses was scored as missing and one as non-response (out of 72 responses across six conditions). Taken together, these data suggest that approximately 6–7% of startle responses may need to be excluded from data analyses due to artifacts (which typically occur at random throughout a task). It may therefore be necessary to increase the size of one's task by this percentage so as to improve the likelihood that there are six responses available for analyses.

Although the overarching goal of this study was to provide an empirically determined guideline to inform the development of startle tasks, a second and related goal is to inform data pre-processing and analytic procedures for emotion-modulated startle paradigms. For example, if some participants have multiple unusable trials due to randomly occurring artifacts, researchers may choose to include those participants in analyses so long as there are still six usable trials per condition. Researchers should, however, consider their sample size when determining whether subjects with noisier EMG data should be excluded from analyses. When sample sizes are small, researchers may choose to include subjects with fewer than six usable trials per condition in order to improve the signal-to-noise ratio of the data. Ultimately, research of this nature can also inform the selection of artifact-rejection procedures that strike an appropriate balance to maximize signal-to-noise ratio. Two important caveats to the abovementioned guideline (i.e., the minimum NoS per condition = six) should be noted. First, this guideline may only generalize to studies that utilize the NPU-threat task (Schmitz and Grillon, 2012). That is, a different NoS may be (and likely will be) necessary to obtain reliable signals for other emotion modulated startle paradigms (e.g., affective picture viewing [e.g., Lang et al., 1997], or fear conditioning [Duits et al., 2015]). Second, given that the present study's clinical sample only included individuals with select internalizing disorders (i.e., MDD and/or PD), the suggested minimum NoS may not apply to individuals with other types of psychopathologies, such as externalizing or psychotic disorders.

There are also several limitations to the present study that should be noted. First, the two samples had slightly different NPU-threat tasks (e.g., countdowns vs. geometric shapes for cues), although the overall recommended NoS for both samples were quite comparable. Second, the sample size for retest reliability analyses was too small to evaluate whether retest reliability differed by diagnosis. Third, although analyses were also conducted with startle responses added in a random order (see Footnote 1), startle responses were only randomized once for this purpose. Thus, future studies should examine whether results change as a function of repeated random sampling. Additionally, further studies should examine whether a similar NoS is necessary to obtain a reliable index of baseline startle magnitude. However, this study benefited from several strengths including the assessment of the reliability of startle across two samples, one of which included individuals with diagnosed internalizing psychopathology. Additionally, the reliability of startle magnitude *and* amplitude were examined, which is important given that these two methods of startle quantification are each frequently used in research.

5. Conclusions

Results from the present study provide information that may help researchers obtain psychometrically sound indices of emotional processing using the eyeblink startle reflex. In particular, our findings suggest that a minimum of *six* responses may be sufficient for obtaining a reliable and stable index of emotion-modulated startle (i.e., anxiety-potentiated and fear-potentiated startle) during the NPU-threat task in non-clinical and clinical samples. Although this guideline may apply to

other emotion-modulated paradigms, future studies should test this directly.

References

Abd-El-Fattah, S.M., Hassan, H.K., 2011. Dependent-alpha calculator: testing the differences between dependent coefficients alpha. Journal of Applied Quantitative Methods 6, 59–61.

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. http://dx.doi.org/10.1111/j.1469-8986.2005.00271.x.

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. http://dx.doi.org/10.1111/psyp.12545.

Bradley, M.M., Gianaros, P., Lang, P.J., 1995. As time goes by: stability of affective startle modulation. Psychophysiology 32, 21.

Campbell, M.L., Gorka, S.M., McGowan, S.K., Nelson, B.D., Sarapas, C., Katz, A.C., ... Shankman, S.A., 2014. Does anxiety sensitivity correlate with startle habituation? An examination in two independent samples. Cognit. Emot. 28 (1):46–58. http://dx.doi.org/10.1080/02699931.2013.799062.

Carleton, R.N., Norton, M.P.J., Asmundson, G.J., 2007. Fearing the unknown: a short version of the Intolerance of Uncertainty Scale. J. Anxiety Disord. 21 (1):105–117. http://dx.doi.org/10.1016/j.janxdis.2006.03.014.

Cornwell, B.R., Johnson, L., Berardi, L., Grillon, C., 2006. Anticipation of public speaking in virtual reality reveals a relationship between trait social anxiety and startle reactivity. Biol. Psychiatry 59 (7):664–666. http://dx.doi.org/10.1016/j.biopsych.2005.09.015.

Craske, M.G., Rauch, S.L., Ursano, R., Prenoveau, J., Pine, D.S., Zinbarg, R.E., 2009. What is anxiety disorder? Depress. Anxiety 26 (12):1066–1085. http://dx.doi.org/10.1002/da.20633.

Cronbach, L.J., 1947. Test "reliability": its meaning and determination. Psychometrika 12 (1):1–16. http://dx.doi.org/10.1007/BF02289289.

Cronbach, L.J., Meehl, P.E., 1955. Construct validity in psychological tests. Psychol. Bull. 52 (4):281–302. http://dx.doi.org/10.1037/h0040957.

Duits, P., Cath, D.C., Lissek, S., Hox, J.J., Hamm, A.O., Engelhard, I.M., ... Baas, J.M., 2015. Updated meta-analysis of classical fear conditioning in the anxiety disorders. Depress. Anxiety 32 (4):239–253. http://dx.doi.org/10.1002/da.22353.

First, M.B., Spitzer, R.L., Gibbon, M., Williams, J.B.W., 1996. Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I). Biometric Research Department, New York.

Foti, D., Kotov, R., Hajcak, G., 2013. Psychometric considerations in using error-related brain activity as a biomarker in psychotic disorders. J. Abnorm. Psychol. 122 (2): 520–531. http://dx.doi.org/10.1037/a0032618.

Gorka, S.M., Lieberman, L., Shankman, S.A., Phan, K.L., 2017. Startle potentiation to uncertain threat as a psychophysiological indicator of fear-based psychopathology: an examination across multiple internalizing disorders. J. Abnorm. Psychol. 126 (1):8–18. http://dx.doi.org/10.1037/abn0000233.

Grillon, C., Ameli, R., 2001. Conditioned inhibition of fear-potentiated startle and skin conductance in humans. Psychophysiology 38 (5):807–815. http://dx.doi.org/10.1017/S0048577201000294.

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. Am. J. Psychiatr. 165:898–904. http://dx.doi.org/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. Biol. Psychiatry 66 (1):47–53. http://dx.doi.org/10.1016/j.biopsych.2008.12.028.

Hajcak, G., Patrick, C.J., 2015. Situating psychophysiological science within the research domain criteria (RDoC) framework. Int. J. Psychophysiol. 98 (2):223–226. http://dx.doi.org/10.1016/j.ijpsycho.2015.11.001.

Kappenman, E.S., Farrens, J.L., Luck, S.J., Proudfit, G.H., 2014. Behavioral and ERP measures of attentional bias to threat in the dot-probe task: poor reliability and lack of correlation with anxiety. Front. Psychol. 5 (9). http://dx.doi.org/10.3389/fpsyg.2014.01368.

Kaye, J.T., Bradford, D.E., Curtin, J.J., 2016. Psychometric properties of startle and corrugator response in NPU, affective picture viewing, and resting state tasks. Psychophysiology http://dx.doi.org/10.1111/psyp.12663.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 1997. Motivated attention: affect, activation, and action. In: Lang, P.J., Simons, R.F., Balaban, M.T. (Eds.), Attention and Orienting: Sensory and Motivational Processes. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 97–135.

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. http://dx.doi.org/10.1111/1469-8986.3710092.

Lieberman, L., Gorka, S.M., Shankman, S.A., Phan, K.L., 2017. Impact of panic on psychophysiological and neural reactivity to unpredictable threat in depression and anxiety. Clin. Psychol. Sci. 5 (1):52–63. http://dx.doi.org/10.1177/2167702616666507.

Lissek, S., Pine, D.S., Grillon, C., 2006. The strong situation: a potential impediment to studying the psychobiology and pharmacology of anxiety disorders. Biol. Psychol. 72 (3), 265–270.

Manber, R., Allen, J.J.B., Burton, K., Kaszniak, A.W., 2000. Valence-dependent modulation of psychophysiological measures: is there consistency across repeated testing? Psychophysiology 37 (5):683–692. http://dx.doi.org/10.1111/1469-8986.3750683.

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

Moran, T.P., Jendrusina, A.A., Moser, J.S., 2013. The psychometric properties of the late positive potential during emotion processing and regulation. Brain Res. 1516: 66–75. http://dx.doi.org/10.1016/j.brainres.2013.04.018.

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. http://dx.doi.org/10.1111/psyp.12418.

Nunnally, J.C., 1978. Psychometric Theory. second ed. McGraw-Hill, New York.

Santos, J., 1999. Cronbach's alpha: a tool for assessing the reliability of scales. J. Ext. 37 (2), 34–36.

Sarapas, C., Weinberg, A., Langenecker, S.A., Shankman, S.A., 2017. Relationships among attention networks and physiological responding to threat. Brain Cogn. 111, 63–72.

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

Schwartz, S.J., Lilienfeld, S.O., Meca, A., Sauvigné, K.C., 2016. The role of neuroscience within psychology: a call for inclusiveness over exclusiveness. Am. Psychol. 71 (1):52–70. http://dx.doi.org/10.1037/a0039678.

Shankman, S.A., Gorka, S.M., 2015. Psychopathology research in the RDoC era: unanswered questions and the importance of the psychophysiological unit of analysis. Int. J. Psychophysiol. 98 (2), 330–337.

Shankman, S.A., Nelson, B.D., Sarapas, C., Robison-Andrew, E., 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. J. Abnorm. Psychol. 122 (2):322–338. http://dx.doi.org/10.1037/a0030747.

Tavakol, M., Dennick, R., 2011. Making sense of Cronbach's alpha. International Journal of Medical Education 2:53–55. http://dx.doi.org/10.5116/ijme.4dfb.8dfd.

Tomarken, A.J., 1995. A psychometric perspective on psychophysiological measures. Psychol. Assess. 7 (3):387–395. http://dx.doi.org/10.1037/1040-3590.7.3.387.

Vrana, S.R., Spence, E.L., Lang, P.J., 1988. The startle probe response: a new measure of emotion? J. Abnorm. Psychol. 97 (4):487–491. http://dx.doi.org/10.1037/0021-843X.97.4.487.

Watson, D., O'Hara, M.W., Simms, L.J., Kotov, R., Chmielewski, M., McDade-Montez, E.A., ... Stuart, S., 2007. Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). Psychol. Assess. 19 (3):253. http://dx.doi.org/10.1037/1040-3590.19.3.253.