Neurobiology of Stress 8 (2018) 134–146

Contents lists available at ScienceDirect

Neurobiology of Stress

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

Roadmap for optimizing the clinical utility of emotional stress paradigms in human neuroimaging research

Timothy J. McDermott[a,b], Namik Kirlic[a], Robin L. Aupperle[a,c,*]

[a] Laureate Institute for Brain Research, Tulsa, OK, United States
[b] Department of Psychology, University of Tulsa, Tulsa, OK, United States
[c] Department of Community Medicine, University of Tulsa, Tulsa, OK, United States

ARTICLE INFO

Keywords:
fMRI
Validity
Reliability
Treatment
Fear
PTSD

ABSTRACT

The emotional stress response is relevant to a number of psychiatric disorders, including posttraumatic stress disorder (PTSD) in particular. Research using neuroimaging methods such as functional magnetic resonance imaging (fMRI) to probe stress-related neural processing have provided some insights into psychiatric disorders. Treatment providers and individual patients would benefit from clinically useful fMRI paradigms that provide information about patients' current brain state and responses to stress in order to inform the treatment selection process. However, neuroimaging has not yet made a meaningful impact on real-world clinical practice. This lack of clinical utility may be related to a number of basic psychometric properties that are often overlooked during fMRI task development. The goals of the current review are to discuss important methodological considerations for current human fMRI stress-related paradigms and to provide a roadmap for developing methodologically sound and clinically useful paradigms. This would include establishing various aspects of reliability, including internal consistency, test-retest and multi-site, as well as validity, including face, content, construct, and criterion. In addition, the establishment of standardized normative data from a large sample of participants would support our understanding of how any one individual compares to the general population. Addressing these methodological gaps will likely have a powerful effect on improving the replicability of findings and optimize our chances for improving real-world clinical outcomes.

1. Introduction

Research using functional magnetic resonance imaging (fMRI) has made remarkable achievements that have shaped our current understanding of the neural mechanisms underlying emotional stress and their contributions to psychiatric disorders. While much has been accomplished within this field, there is still room for improvement, particularly in regard to clinical utility. Treatment providers and individual patients would benefit from clinically useful fMRI paradigms that help to provide information about how patients' current brain state is related to their symptoms and responses to stress, as well as to inform the treatment selection process. This information could be particularly useful as a psychoeducational tool to inform patients about how treatments target specific neural mechanisms. The goals of the current review are to discuss important methodological considerations for current human fMRI stress-related paradigms and to provide a roadmap for developing methodologically sound and clinically useful paradigms. While this review focuses primarily on the application of fMRI to studying the neurophysiology underlying emotional stress, many of the recommendations discussed in this review are applicable to other task-based neuroimaging modalities and other subfields of cognitive and clinical neuroscience.

There have been numerous studies using fMRI to examine the neurobiological underpinnings of emotional stress, and these studies have yielded a wealth of important information about the nature of emotional stress in healthy individuals and those with psychiatric disorders (for reviews see: Hughes and Shin, 2011; Shin and Liberzon, 2010). While these studies have been useful in discovering neural markers of psychiatric disorders, the impact of this research on clinical practice has been astoundingly lacking (Carter et al., 2008; Paulus, 2015). While dissemination of research into the clinical realm is an obstacle across mental health research (McClean and Foa, 2013; Shafran et al., 2009; Southam-Gerow et al., 2012), it is often difficult to even imagine specific ways in which our current fMRI findings could be disseminated clinically at this point in time. This is despite the fact that the stress response is considered important in the development and maintenance of so many different mental health disorders (Riboni and Belzung, 2017; Zorn et al., 2017). The lack of clinical utility may in part

[*] Corresponding author. Laureate Institute for Brain Research, 6655 S Yale Ave, Tulsa, OK 74136, United States.
E-mail address: raupperle@laureateinstitute.org (R.L. Aupperle).

https://doi.org/10.1016/j.ynstr.2018.05.001
Received 6 November 2017; Received in revised form 30 April 2018; Accepted 3 May 2018
Available online 07 May 2018

be due to the currently accepted methods for developing and testing fMRI tasks that turn a blind eye to potential future use in clinical settings with individual patients. In comparison to the methodical and step-wise process of developing a psychological assessment that is typical within the fields of neuropsychology or clinical self-report, a number of basic psychometric properties tend to be overlooked during fMRI task development. This includes considerations essential to reliability and validity (Groth-Marnat, 2009). At its most basic, the process of developing an assessment would include establishing reliability, meaning that the measure is stable and consistent, as well as validity, meaning that the assessment actually is measuring what it is designed to measure. In addition, the establishment of standardized normative data from a large sample of participants would support our understanding of how any one individual compares to the general population. While the cost of fMRI is one notable obstacle for its use as a method for clinical assessment, this is not insurmountable in all cases – particularly given the widespread use of MRI scanners in hospital settings. Notably, other fields of medicine routinely utilize relatively expensive tests (e.g., anatomical MRIs, CT scans, EKGs, EEGs, etc.), reasoning that it is preventative of future medical problems and thus reduces future costs. In addition, developing fMRI paradigms that have known clinical utility would then allow for identification of cheaper methods for estimating neural stress response patterns (i.e., using EEG, physiological, or behavioral indices). Lastly, even if fMRI is only to be used as a tool for identifying biomarkers related to current and novel treatment targets in research (rather than clinical settings), psychometric issues remain a central obstacle.

There are several caveats and complexities to human neuroimaging research that have perhaps hindered the development of psychometrically-sound paradigms. Perhaps most notably, human clinical neuroscience remains an emerging field of study, and the process of establishing standards for the field is still underway. In addition to standardizing the process by which fMRI paradigms are developed, there is also a need for standardizing the preprocessing and analysis methods used when analyzing neuroimaging data acquired using fMRI. This is especially important since there have been a number of recent concerns about the reproducibility and generalizability of findings from neuroimaging studies (Poldrack et al., 2016). Most human neuroimaging research to date takes an approach that is exploratory and requires a certain degree of fine-tuning depending on the population being studied and the paradigm being used. This malleable approach to data analysis, along with a certain lack of transparency in procedure, presents another potential roadblock to future clinical utility.

A clinically useful fMRI paradigm related to emotional stress would be one that would allow clinicians to better understand how a specific patient responds to and processes stress and in turn, make judgments about what treatments might be most appropriate and hopefully lead to better outcomes for that patient. In this review, we attempt to lay out a roadmap for developing fMRI assessments that elucidate the emotional stress response in ways that could be useful in a clinical setting. To facilitate the use of concrete examples demonstrating the use of such a ‘clinical-utility roadmap,’ we will focus specifically on posttraumatic stress disorder (PTSD) as a clinical group for whom the neurobiological stress response is highly relevant. We start with a review regarding the various types of fMRI paradigms used to identify neural substrates of emotional stress, including summaries of results concerning PTSD. We then summarize research with current stress-related fMRI paradigms in regards to reliability, validity, or the development of normative data. Finally, we outline a roadmap regarding optimal steps for supporting the further development of clinically useful emotional stress paradigms.

2. Paradigms used to assess neural responses to emotional stress

A number of paradigms have been essential to our understanding of stress- and fear-related neurocircuitry and associated psychiatric disorders. These paradigms can be grouped into two domains, the *cognitive-behavioral* and *symptom provocation* paradigms (Rauch et al., 2003). Cognitive-behavioral paradigms utilize tasks designed to engage specific brain systems of interest, such as (1) passive viewing of facial stimuli, (2) passive viewing of emotionally valenced pictures, and (3) classical fear conditioning. Conversely, in symptom provocation paradigms, response to intentionally induced, disorder-relevant symptoms and control conditions are measured. Both sets of paradigms lend themselves to examinations across patient and healthy control populations, behavioral and pharmacological manipulations, as well as longitudinal and treatment studies. While there have been tasks developed to specifically probe various aspects of emotion regulation (i.e., cognitive reappraisal tasks; Ochsner et al., 2002; Rabinak et al., 2014), we restrict our following research summary to paradigms that have been more extensively utilized in research with PTSD, namely emotional face processing, emotional images/scenes, fear conditioning, and symptom provocation.

2.1. *Emotional faces*

Paradigms employing facial stimuli usually have subjects passively view presentations of faces showing various types of emotions (e.g., fearful, sad, angry, disgusted, and happy), as well as neutral faces. Early studies compared responses between differently valenced facial stimuli, such as fearful vs. neutral or angry vs. happy faces (Rauch et al., 2003; Davis and Whalen, 2001). Another early study compared responses between matching similarly valenced facial stimuli and matching shapes (Hariri et al., 2002). Backward masking has also been used to assess face processing below conscious awareness (e.g., Whalen et al., 1998). Recent variations of emotional face paradigms require subjects to identify the gender of the presented face (Blair et al., 2008), label the salient affect/emotion (Fonzo et al., 2017), or match the face with available affective words (Robinson et al., 2012).

Studies show that passive viewing of human face stimuli robustly produces amygdala activation (Fig. 1; Phan et al., 2002). While the amygdala is found to activate to facial stimuli regardless of valence (Breiter et al., 1996), there is some indication that it does so preferentially for fearful faces (e.g., Whalen et al., 2001; Whalen et al., 1998). It also appears that the amygdala is particularly sensitive to processing of facial emotional stimuli, as the amygdala has been found to activate more for emotional facial expression stimuli than emotional images (Hariri et al., 2002). Other regions activated by emotional faces include parahippocampal gyrus (PHG), posterior cingulate cortex (PCC), middle temporal gyrus, insula, ventromedial prefrontal cortex (vmPFC), dorsal anterior cingulate cortex (dACC), orbitofrontal cortex (OFC), visual cortex, putamen, and the cerebellum (Fig. 1; Fusar-Poli et al., 2009; Posamentier and Abdi, 2003).

2.2. *Emotional images and scenes*

The International Affective Picture System (IAPS; Lang et al., 2008) has been used widely to study emotional processing. In a typical experiment, subjects passively view pleasantly and unpleasantly valenced pictures of varying arousal level expected to elicit corresponding emotions or stress responses (Fig. 1). Unpleasant, or aversive pictures for example include scenes of motor vehicle accidents, violence, or mutilated bodies, while pleasant pictures include images of families, animals, or neutral objects. When these paradigms are used with psychiatric populations, the content of the unpleasant or aversive pictures is not disorder-specific. Rather, these paradigms seek to probe emotional processing more generally (in contrast to symptom provocation paradigms, as described below). Researchers often examine both the anticipation of and responses to emotional images (Nitschke et al., 2006; Simmons et al., 2006). Emotional pictures seem to activate a similar neural network as emotional face paradigms, including the amygdala, PHG, PCC, insula, vmPFC, dACC, OFC, and visual cortex (Fig. 1; Phan et al., 2002; Britton et al., 2005).

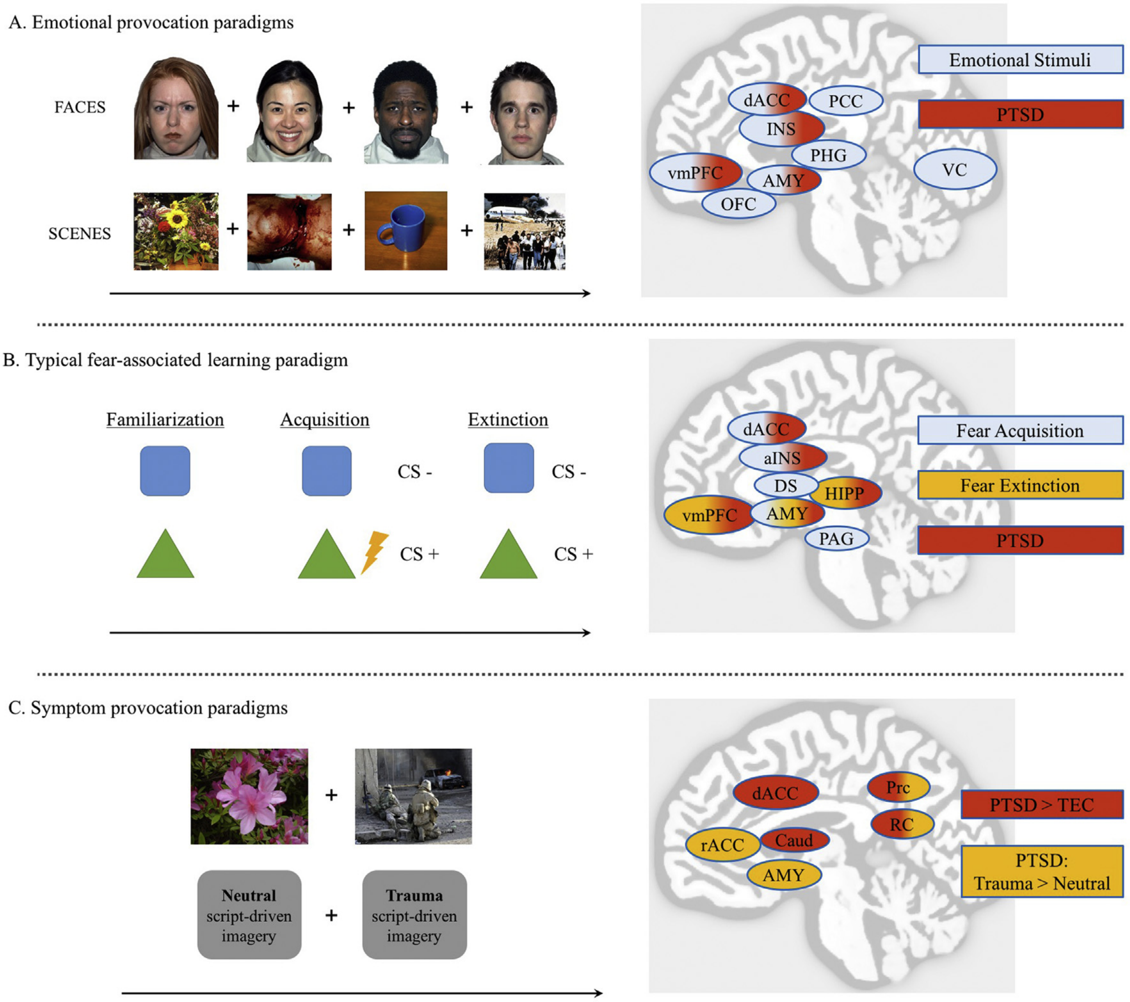

Fig. 1. Overview of Emotional Stress Paradigms.
A. Emotional provocation paradigms. Left –Examples of stimuli used in emotional faces (Tottenham et al., 2009) and emotional scenes (Lang et al., 2008) paradigms. Right – Network of brain regions activated during the presentation of emotional stimuli (blue) and implicated in PTSD (red).
B. Typical fear-associated learning paradigm. Left – Examples of conditioned stimuli (CS) used in fear-associated learning paradigms. The CS+ (green triangle) is associated with an aversive shock (unconditioned stimulus; US) during fear acquisition (marked by electricity symbol), while the CS- (blue square) is not associated with the US. Right – Network of brain regions activated during fear acquisition (blue), fear extinction (yellow), and implicated in PTSD (red).
C. Symptom provocation paradigms. Left – Examples of stimuli, both neutral and trauma-related, used in script-driven symptom provocation paradigms. Right – Network of brain regions in which patients with PTSD exhibit significantly greater activation (red) than trauma-exposed controls (TEC) or where patients with PTSD exhibit significantly greater activation to trauma-related stimuli relative to neutral stimuli (yellow).
Abbreviations: anterior cingulate cortex (ACC), dorsal ACC (dACC), rostral ACC (rACC), caudate nucleus (Caud), precuneus (Prc), retrosplenial cortex (RC), amygdala (AMY), parahippocampal gyrus (PHG), posterior cingulate cortex (PCC), prefrontal cortex (PFC), ventromedial PFC (vmPFC), orbitofrontal cortex (OFC), insula (INS), anterior insula (aINS), visual cortex (VC). hippocampus (HIPP), periaqueductal gray matter (PAG), dorsal striatum (DS). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

2.3. *Fear conditioning*

Conditioning refers to the process of learning the association between two previously unrelated stimuli (Pavlov, 1927). In a typical fear (Pavlovian) conditioning paradigm, a previously neutral conditioned stimulus (CS) is paired with an aversive fear-inducing unconditioned stimulus (US) (Fanselow and Ponnusamy, 2008). After repeated presentations, the CS alone elicits a conditioned fear response (CR) that occurs independently of the US. Conversely, repeated presentations of the CS without the US gradually result in weakening of the CR, a process called extinction. While fear conditioning can examine how fear is learned, extinction focuses on how fear is extinguished, or rather how safety is learned and retained.

Although very similar to animal and non-imaging studies, human fear conditioning and extinction neuroimaging protocols have employed a tone, electric shock, thermal stimulus, air blast, odor, or an aversive image as US, and geometric visual figures, videos, light, tones, or faces as CS. Protocols vary on whether they employ trace (CS is presented some brief time before the US) or delayed conditioning (CS and US overlap, or the CS is immediately followed by the US), as well as on rate of reinforcement. While the US is typically brief in duration, the length of CS can range from brief to sustained, which is useful when fear vs. anxiety (i.e., anticipatory) responses are of interest (Schmitz and Grillon, 2012). In addition to the US-reinforced CS (CS+), human neuroimaging studies also typically employ a second, non-reinforced CS (CS-), which is not associated with the US (Fig. 1). This allows researchers to identify unique fear conditioned neural and psychophysiological responses to the CS+. Finally, protocols vary on retention (CRs in environment where extinction was learned), renewal (CRs in environment where fear was learned), and reinstatement (CRs to US after extinction is established) are tested (Lonsdorf et al., 2017). Related paradigms also include observational fear conditioning paradigms (i.e., social learning of fear; Olsson and Phelps, 2007; Hygge and Ohman, 1978; Haaker et al., 2017) and instructed threat paradigms (where subjects are verbally instructed about the CS-US relationship; Mechias et al., 2010).

Several reviews and meta-analyses have summarized the neurocircuitry of fear conditioning and extinction (VanElzakker et al., 2014; Fullana et al., 2015; Liberzon and Abelson, 2016; Hughes and Shin, 2011; Shin and Liberzon, 2010). In brief, this extensive body of research suggests that fear conditioning and extinction activate a functional network that includes amygdala, hippocampus, prefrontal cortex, anterior insula, dorsal striatum and midbrain structures (Fig. 1). During fear learning and expression, the amygdala is thought to play a crucial role in detection of salient stimuli and thus also in fear learning and expression, while the hippocampus is considered important for processing of information related to threat cues and contexts. Prefrontal regions, including dACC, dorsomedial prefrontal cortex (dmPFC), and vmPFC have been implicated in conscious threat appraisal, and particularly during anticipatory threat (Fig. 1). The anterior insula is believed to integrate cognitive, affective, and interoceptive states during conditioning to create a subjective representation of a fear experience. Fear expression (i.e., physiological autonomic arousal, defensive responses, avoidance behavior) is thought to be further modulated by the dorsal striatum, hypothalamus and the periaqueductal gray matter. In regard to extinction learning and recall, human neuroimaging studies have at least partially supported animal work by highlighting the importance of vmPFC-amygdala-hippocampus networks.

2.4. Symptom provocation

Symptom provocation paradigms utilize disorder-related (e.g., trauma stimuli for PTSD [Hendler et al., 2003], phobic stimuli for specific phobia [Schienle et al., 2007], social evaluation for social anxiety [Boehme et al., 2014]) reminders such as pictures or script-driven imagery to uncover neuronal responses. Script-driven imagery procedures typically involve subjects composing several individualized scripts depicting personally experienced traumatic, neutral, and positive events, as well as various hypothetical events (Lanius et al., 2001; Hopper et al., 2007). While symptom provocation paradigms most often utilize visual stimuli, paradigms can also utilize other stimuli to provoke symptoms (e.g., anticipatory threat of aversive electrical stimulation to provoke anxiety symptoms [Schunck et al., 2008]). Findings from symptom provocation studies across anxiety- or trauma-related disorders often implicate common prefrontal-insula-amygdala networks (Etkin and Wager, 2007). Below, we have summarized neuroimaging findings relevant to PTSD specifically.

2.5. PTSD findings

The above described cognitive-behavioral paradigms have been used to assess differences between individuals with PTSD and healthy controls. A number of reviews and meta-analyses have summarized the neurocircuitry of PTSD (Fig. 1; Etkin and Wager, 2007; Hughes and Shin, 2011; Shin and Liberzon, 2010; VanElzakker et al., 2014). Findings reflect relative functional alterations in PTSD patients in a number of brain regions involved in emotional processing and regulation, including the amygdala, hippocampus, insula, dACC, and vmPFC. Interpretations of these data often center on the idea that amygdala and insula regions are hyperresponsive in PTSD, while portions of the vmPFC may be hypoactive, failing to inhibit the amygdala. However, directionality of findings within these regions are often inconsistent, perhaps due to differences in experimental paradigms utilized or specific trauma populations included. Nevertheless, these alterations are thought to contribute to the exaggerated fear responses and persistence of traumatic memories, as well as deficits in extinction, emotion regulation, attention, and contextual processing observed in PTSD. Particularly clinically relevant for our understanding of PTSD are symptom provocation paradigms. A meta-analysis of 19 such studies showed that patients with PTSD exhibit significantly greater activation in retrosplenial cortex, precuneus, rostral ACC, and bilateral amygdala in response to trauma-relevant compared to neutral stimuli (Fig. 1; Sartory et al., 2013). These studies provide evidence for PTSD-related alterations in regions not only involved in salience-detection, but also self-referential and autobiographical processes.

As the literature delineating neural networks related to emotional stress in PTSD (and other psychiatric disorders) has been growing, there has been increasing attention towards how we might use such information to improve our clinical treatments (Milad et al., 2014; Paulus, 2015). However, it is crucial to also recognize the methodological research that has been conducted on these stress-related fMRI paradigms, as it has important implications for how neuroimaging assessments may or may not be useful in the clinical realm.

3. Reliability of emotional stress paradigms

To improve the clinical utility of fMRI emotional stress paradigms, it is necessary to demonstrate their reliability. If the neural activity measured by an fMRI scan changes significantly based on when or where a scan is completed, clinicians cannot make meaningful recommendations to patients regarding their brain state and its relation to symptomatology. The primary forms of reliability are test-retest reliability, inter-rater reliability, and internal consistency. In the sections below, we will discuss each of these forms of reliability in the context of human neuroimaging and review previous fMRI reliability studies using emotional stress paradigms.

3.1. Test-retest reliability

Test-retest reliability refers to the stability of an assessment in its measurements across time. In regard to clinical utility, it is necessary to demonstrate that the measured brain activity is representative of more than just a single time point. For example, a patient with PTSD may show an altered amygdala response during a stress-related paradigm, but unless this alteration persists over time (at least without intervention or significant change in mental state), any relevant recommendations would be misguided. Test-retest reliability is also necessary for making meaningful conclusions in treatment studies. For example, a study might be examining whether a specific intervention may be useful for targeting specific patterns of the neurophysiological stress response. In order for the findings of this study to provide meaningful information about the impact of the intervention, it would have to first be established that a person who completes the same task at two different time points without any sort of intervention has a similar neurophysiological response profile. While treatment studies may include a control group to help account for such repeated measurements, this is not sufficient to account for potential errors in measurement. Rather, to support optimal clinical utility, it must be established that the response profile of an *individual person* is reliable across time.

To establish test-retest reliability, researchers typically conduct studies where participants complete the same fMRI protocol on different days and then examine how well neuronal activity correlates between the different scanning sessions. The intraclass correlation coefficient (ICC) provides a quantitative value of the consistency of multiple measurements (Shrout and Fleiss, 1979), and it is a commonly used statistical method utilized in test-retest reliability studies (Bennett and Miller, 2010; Weir, 2005; Yen and Lo, 2002) As compared to Pearson's r correlation coefficients, ICCS are better able to detect systematic error to determine absolute agreement between multiple measurements (Weir, 2005). There are several different types of ICCs that can be calculated in test-retest reliability studies, and we will focus on the absolute ICC [i.e., ICC(2, 1)] and relative ICC [i.e., ICC(3, 1)]. The absolute ICC looks at the reliability of a measure by comparing exact agreement of data across subjects and sessions of a variable (Shrout and

Fleiss, 1979). This means that a high absolute ICC indicates that a specific individual is likely to have a similar value at test and retest, regardless of the comparison group. Relative ICC looks at the consistent reliability using the means of a variable across sessions, meaning that the calculation of each individual's measurement is relative to the measurements obtained from others (Shavelson and Webb, 1991; Shrout and Fleiss, 1979). Relative ICCs tend to have higher values than absolute ICCs since systematic changes in the mean are statistically controlled for, and thus, relative ICC is more comparable to Pearson's r than absolute ICC (Shavelson and Webb, 1991). Each ICC is also calculated using either a "single-measure" or an "average-measure" approach (Shavelson and Webb, 1991). Most studies utilize either relative or absolute "single-measure" ICCs, as these are meant to compare the agreement of independent measurements (Shavelson and Webb, 1991). Average-measure ICCs, which could also be absolute or relative, consider the consistency across multiple measurements that will eventually be averaged together (Shavelson and Webb, 1991) (for example, in the case of internal consistency or some cases of inter-rater reliability).

When reporting findings from previous test-retest studies of emotional faces tasks in fMRI, we indicate the type of ICC calculation that was used. The interpretation of ICC values typically follows the guidelines presented in Fleiss (1986), which are that ICCs less than 0.4 are considered poor, ICCs between 0.40 and 0.59 are considered fair, ICCs between 0.60 and 0.74 are considered good, and ICCs greater than or equal to 0.75 are considered excellent. Note that there can be cases where ICCs are negative. ICCs are calculated using ratios of between-subject and within-subject variance, and negative ICCs occur when the within-subject variance exceeds the between-subject variance (Lahey et al., 1983). Negative ICCs should be interpreted as having a reliability of zero (Bartko, 1976).

An equally important methodological consideration for fMRI test-retest reliability studies is the process of determining which neurophysiological measurements are actually used to calculate ICCs. Most studies tend to use regions-of-interest (ROIs) that have previously been shown to be activated during a specific emotional stress paradigm. ROIs can be defined anatomically by using all of the voxels from within an entire region derived from a brain atlas, or ROIs can be defined functionally by using only the voxels that are significantly active during task performance. Once the ROIs are defined, mean contrast values across all voxels in each ROI from an individual are averaged and then are used to calculate ICCs. Other approaches use procedures for selecting individual voxels (i.e., using the peak activation voxel within a region) to calculate ICCs. The implications of these and other methodological variations will be discussed in regard to the findings from individual studies. Below, we review the available studies for each type of stress-relevant task reviewed in the previous sections, namely *cognitive-behavioral* (includes emotional faces, emotional images/scenes, and fear conditioning) and *symptom provocation* tasks.

3.1.1. Emotional faces

There have been several test-retest reliability studies using facial stimuli paradigms (Johnstone et al., 2005; Schacher et al., 2006; Manuck et al., 2007; Plichta et al., 2012, 2014; Sauder et al., 2013; van den Bulk et al., 2013; Lipp et al., 2014; Nord et al., 2017), the results of which are summarized in Table 1 (including information about tasks, test-retest intervals, contrasts, ROIs, and type and value of ICCs). These studies have primarily focused on the amygdala, though as shown in Table 1, other regions have also been reported. The ICCs reported have varied widely, even within the amygdala. This is likely due to different methods used in regard to the specific contrast or how regions of interest are defined. Specifically, ICCs tended to be greater when contrasting face trials (regardless of valence) to either shape trials or fixation as compared to ICCs when contrasting between specific face valences (e.g., fearful versus happy; Johnstone et al., 2005; Sauder et al., 2013; Lipp et al., 2014). In regard to regions of interest, analyses utilizing functionally-defined ROIs (Johnstone et al., 2005; Schacher et al., 2006; Manuck et al., 2007; Sauder et al., 2013; van den Bulk et al., 2013) typically led to higher ICCs than analyses utilizing anatomically-defined ROIs (Johnstone et al., 2005; Sauder et al., 2013; van den Bulk et al., 2013; Lipp et al., 2014). However, studies using peak voxels showed similar reliability to anatomically-defined ROIs using the same data (Plichta et al., 2012; Nord et al., 2017).

Of particular interest, three test-retest studies used the same dataset but analyzed the data in different ways (Cao et al., 2014; Plichta et al., 2012, 2014). The 2012 study found poor reliability for mean activation in the amygdala both when using anatomically-defined ROIs and peak voxels (all ICCs < 0.18; Table 1). Rather than calculating reliability for mean activation in the amygdala, the Plichta et al., 2014, study instead calculated reliability of amygdala habituation during task performance. The habituation of the amygdala in response to threatening faces has been noted in the literature for some time (Breiter et al., 1996; Wright et al., 2001; Fischer et al., 2003), but this was the first paper to directly examine how this relates to reliability. Amygdala habituation was calculated using two different methods: (1) the amplitude difference between the first and last of the four presentation blocks and (2) modeling of habituation using a regression approach that estimated the rate of habituation (Plichta et al., 2014). They found improved ICCs for both the left and right amygdala using both habituation calculations (most ICCs in the 0.40 - 0.50 range; Table 1). The findings from this study demonstrated that amygdala habituation was more reliable across time than mean activation amplitude. Cao et al. (2014) also calculated reliability using an non-traditional methodology. In this case, they used principles of graph theory, which consider the way that activity from multiple regions correlate when simultaneously active (Cao et al., 2014). Compared to their original study, the ICCs were improved when using global connectivity (ranging from 0.51 to 0.68). The findings from these two studies are important in that they demonstrate how reliability may be improved when using alternative methodologies for analysis. However, it is also important to consider how the measures extracted using these methodologies are different from previous studies comparing healthy and clinical populations. Thus, the clinical relevance of these methods would also have to be determined (see validity section below).

3.1.2. Emotional images and scenes

There has only been a single fMRI test-retest reliability study using an emotional images/scenes task. Stark et al. (2004) conducted a test-retest reliability study in which twenty-four volunteers viewed emotional scenes in a block design during two sessions, one week apart. They examined the BOLD response to either fear-inducing or disgust-inducing IAPS images contrasted against neutral IAPS images. Cohen's Kappa, which is a reliability metric that is different from the ICC but similarly ranges from 0 to 1, was used to calculate the reliability of significant activations across the whole brain. The statistically active regions included the amygdala, vmPFC, OFC, PCC, and hippocampus. Cohen's Kappas showed low reliability of the activations between the two scans with the median value falling below 0.1. While these results were not encouraging regarding the reliability of activation to emotional images/scenes, it could be useful for future research to examine whether reliability may be improved by utilizing an ROI-based approach or other alternative methodologies.

3.1.3. Fear conditioning and extinction

Thus far, there has not been a single fMRI fear conditioning/extinction study published that focuses on test-retest reliability, though there have been several behavioral test-retest reliability studies (Fredrikson et al., 1993; Zeidan et al., 2012; Torrents-Rodas et al., 2014). All of these studies used skin conductance responses (SCRs) to quantify the fear response, while Torrents-Rodas et al. (2014) also included a measure of startle response. Each study had two testing sessions on different days, utilizing two versions of the task (each with different stimuli) to avoid habituation between sessions. Findings from

Table 1
Test-retest reliability results from studies using emotional faces tasks.

Author, Year	N	Test-retest Interval	Task	Contrast	Method	ROI	ICCs
Johnstone et al., 2005	15	0, 2, 8 weeks (3 scans)	Viewing	Faces–Fixation (Fearful)	Anatomical ROIs	L Amygdala	.28
						R Amygdala	.18
					Functional ROIs	L Amygdala	**.67**
						R Amygdala	**.44**
Schacher et al., 2006	12	Range: 1–8 weeks	Viewing	Faces-Baseline (Fearful)	Functional ROIs	L Amygdala	**.83**
						R Amygdala	**.69**
Manuck et al., 2007	13	Median: 21 months Range: 13–22 months	Matching	Faces–Shapes (Fearful/Angry)	Functional ROIs	L Amygdala	-.08
						R Amygdala	**.59**
Plichta et al., 2012	25	Mean: 14.6 days Range: 12–21 days	Matching	Faces–Shapes (Fearful/Angry)	Anatomical ROIs	L Amygdala	.16
						R Amygdala	–.02
					Peak Voxels	L Amygdala	.18
						R Amygdala	.07
Plichta et al., 2014	25	Mean: 14.6 days Range: 12–21 days	Matching	FmL: Habituation (Fearful/Angry)	Anatomical ROIs	L Amygdala	**.41**
						R Amygdala	**.48**
				Regression: Habituation (Fearful/Angry)	Anatomical ROIs	L Amygdala	.25
						R Amygdala	**.53**
Sauder et al., 2013	27	Mean: 88.9 days	Matching	Faces–Shapes (Fearful)	Anatomical ROIs	L Amygdala	**.42**
						R Amygdala	.36
					Functional ROIs	L Amygdala	.32
						R Amygdala	**.40**
						L FFA	**.50**
						R FFA	**.50**
van den Bulk et al., 2013	18	0, 3, 6 months (3 scans)	Labeling	Faces-Fixation (Fearful/Happy/ Neutral)	Anatomical ROIs	L Amygdala	.28*
						R Amygdala	.34*
					Functional ROIs	L Amygdala	.10*
						R Amygdala	.35*
Lipp et al., 2014	15	Mean: 23 days Range: 15–34 days	Classifying Gender	Faces–Fixation (Fearful)	Anatomical ROIs	L Amygdala	.29*
						R Amygdala	.19*
						L FFA	-.24*
						R FFA	-.10*
Nord et al., 2017	29	Mean: 14.3 days (3 scans)	Matching	Faces–Shapes (Fearful/Angry)	Anatomical ROIs	L Amygdala	**.43***
						R Amygdala	-.14*
						sgACC	.33*
					Peak Voxels	L Amygdala	.30*
						R Amygdala	-.50*
						sgACC	-.13*
						R FFA	**.83***

Unless otherwise noted, ICCs were absolute. *Indicates relative ICC. ICCs > 0.4 are in bold. Interpretation of ICC values: poor (< 0.40), fair (0.41-0.59), good (0.60-0.74), excellent (> 0.75). Negative ICCs are interpreted as having zero reliability. Abbreviations: L = left; R = right; FmL = first block minus last block; ROI = region-of-interest; FFA = fusiform face area; sgACC = subgenual anterior cingulate cortex. Test-retest intervals are reported as they were in their respective studies. Unless otherwise noted, test-retest involved 2 scans.

two of these studies reported SCRs to have fair to good reliability and that reliability was generally better during the fear conditioning phase than the fear extinction phase (Fredrikson et al., 1993; Zeidan et al., 2012). In the third study, both SCRs and the startle response showed poor reliability, but the reliability findings were also generally better during the fear conditioning than the fear extinction phase (Torrents-Rodas et al., 2014). The findings from these studies have provided valuable information regarding the reliability of the physiological components of fear conditioning and extinction tasks. However, it is important for future work to test the reliability of neural activations during the various phases of fear learning.

3.1.4. Symptom provocation

There has only been a single fMRI test-retest reliability study using a symptom provocation paradigm. Schunck et al. (2008) conducted a test-retest reliability study in which fourteen volunteers completed a paradigm involving anticipation of aversive transcutaneous electric nerve stimulation during two sessions, 10 days apart. They used fMRI to examine the neural response to periods of anticipatory threat (cues associated with 50% likelihood of shock) contrasted against periods of rest (no likelihood of shock). Absolute ICCs were computed using functionally-defined ROIs, defined by centering 8-mm spheres on the local maxima of regions that were significantly active at both time points. This approach may in theory lead to higher ICCs, as it focuses analyses to voxels known to be active at both test and retest. However, whether such an approach is more or less generalizable to other groups or individuals is unknown. In this study, ICCs showed good reliability in the left inferior parietal lobe (ICC = 0.66), fair reliability in the right inferior frontal gyrus/anterior insula (ICC = 0.54), and poor reliability in the left inferior frontal gyrus/anterior insula (ICC = 0.33), cingulate/medial frontal gyrus (ICC = 0.25), and right inferior parietal lobe (ICC = −0.06). While these findings were somewhat encouraging, additional research is needed to determine the most reliable regions of interest and contrasts to use in analyses with the various symptom provocation paradigms.

3.2. Inter-rater/multi-site reliability

Inter-rater reliability refers to stability of an assessment in its measurements across different raters/assessors (Groth-Marnat, 2009). In the context of fMRI, reliability across different neuroimaging research sites (i.e., multi-site reliability), using different scanners, could be considered a form of inter-rater reliability. Multi-site reliability is crucial to supporting the future clinical utility of fMRI. Regardless of the neuroimaging center or treatment at which a patient is being assessed, the results need to be consistent in order to make meaningful interpretations and recommendations. There have only been two multi-site reliability studies using fMRI emotional stress paradigms (Brown et al., 2011; Gee et al., 2015), but these studies have provided invaluable information regarding consistency of findings and provide model frameworks for future multi-site reliability studies.

In a study conducted by Brown et al. (2011), eighteen participants completed the same emotional working memory task across four different MRI sites. ICCs were calculated for activation of every voxel of the brain to emotional versus neutral distractors during visual working memory maintenance. Generally speaking, the authors reported that a greater proportion of variance was attributable to the "person" factor than the "site" factor. They also found that as more data from multiple runs (i.e., greater than four runs) was averaged together, ICCs across most regions of the brain fell within the fair to good range (Brown et al., 2011). The findings support both multi-site and test-retest reliability and provide evidence that using more data (i.e., more trials in a task) can improve reliability across place and time. In a study by Gee et al. (2015), eight participants completed an emotional faces task twice on successive days at eight different MRI sites (Hariri et al., 2002; Lieberman et al., 2007). Similar to the Brown et al. (2011) study, they found that a greater proportion of variance was attributable to the person variance than site variance (Gee et al., 2015). Instead of ICCs, they calculated the generalizability coefficient or the 'G-coefficient', which ranges from 0 to 1 and calculates test-retest reliability using the relative difference in means similar to the relative ICC (Brennan, 2001; Shavelson and Webb, 1991; Webb and Shavelson, 2005). Of interest to this review, the authors calculated reliability for both mean activation and habituation of the amygdala. Similar to the findings mentioned above (Plichta et al., 2014), G-coefficients were somewhat greater for amygdala habituation (mean: 0.32; range: 0.00–0.71) than mean activation (mean: 0.22; range: 0.00–0.44). The G-coefficients reported in the original publication were relatively high compared to previous publications. However, as noted in a recently published clarification on the study, these G-coefficients were somewhat misleading in that they were a product of all sixteen sessions in aggregate (Cannon et al., 2018). The ICC values provided in the clarification were mostly in the poor range. Regardless of test-retest reliability, the findings from these two studies both showed that more variance was attributable to person than to site, which is encouraging for multisite-reliability (Brown et al., 2011; Gee et al., 2015).

3.3. Internal consistency

Internal consistency is the reliability of responses to individual items throughout an assessment (Groth-Marnat, 2009). For fMRI, it would be important to demonstrate that the various stimuli used consistently and reliably invoke the expected behavioral response in individuals (e.g., self-reported ratings of valence and arousal) as well as the expected neural response. While this is rarely done when developing fMRI-related tasks, it could maximize the robustness (and eventual reliability) of the task as a whole.

There has not been much work assessing the internal consistency of emotional stress paradigms. However, a recent study used an emotional faces paradigm and assessed the internal consistency of neural responses in a large sample ($N = 139$) of adolescents (Infantolino et al., 2018). The results demonstrated excellent internal consistency for amygdala activation when looking at face and shape contrasts separately, but the internal consistency estimates were poor for faces-shapes contrasts (Infantolino et al., 2018). The generalizability of these findings outside of an adolescent sample is unclear and thus requires further study.

3.4. Reliability summary

With the exception of emotional face paradigms, there is a dearth of research assessing reliability of stress-related neuroimaging paradigms. Overall, findings from even the emotional face paradigms have been highly variable, ranging from excellent (Johnstone et al., 2005; Schacher et al., 2006; Manuck et al., 2007; Sauder et al., 2013; Gee et al., 2015) to poor test-retest reliability of neural activations (van den Bulk et al., 2013; Lipp et al., 2014; Nord et al., 2017). This high variability in fMRI reliability findings is not limited to emotional paradigms. A review by Bennett and Miller (2010) examined test-retest reliability findings across a broader range of fMRI paradigms. Across 15 studies, they found that the mean ICC value was 0.50 and that mean ICCs for each individual study ranged from poor (0.33) to good (0.66). As made apparent across studies using emotional face paradigms, specific regions of interest and/or specific contrasts may prove to be more reliable than others. Future research should continue to assess the reliability of existing emotional stress paradigms (and identify the most reliable variables), make attempts to improve the reliability of these paradigms, and seek to establish reliability of novel paradigms.

In regard to improving the reliability of existing paradigms, there are some additional considerations unique to fMRI research regarding data preprocessing and analysis methods. These considerations would include decisions about voxel-wise thresholding, cluster correction, and correction for motion or physiological noise. Establishing consistency in these methods across studies or demonstrating that findings are robust to variations in these methods would be beneficial. Other information that might be beneficial would include the time of recording, current level of sleepiness, and any other information that could potentially affect the data. As a standard practice, researchers should provide as much detail as reasonably possible when describing their analysis methods. This will not only allow for other researchers to adequately reproduce their study protocols but can also help to elucidate environmental variables that could be controlled in order to optimize reliability.

Finally, it is important to note that producing more reliable fMRI task paradigms/methods will be beneficial to the field of psychiatric neuroimaging as a whole. Assessments with higher reliability have increased power to detect true effects and also require lower sample sizes to do so (Button et al., 2013). There have been important concerns recently raised about the replicability and robustness of fMRI findings (Poldrack et al., 2016). In addition to increasing sample size and conducting replication studies, improvement of reliability could be an important step in addressing these concerns (Thomas et al., 2017; Button et al., 2013).

4. Validity of stress paradigms

Measuring changes in brain activity during the performance of mental tasks can offer a fascinating window into the mechanics of the human mind. However, these measured changes in activity cannot be considered clinically useful until they are shown to have meaningful relationships to observable, disorder-relevant behavior and related theoretical constructs. Demonstrating the validity of an assessment provides support that an assessment is actually measuring what it is designed to measure. There are several forms of validity to examine when developing psychological assessments. We propose that while neuroimaging research has addressed validity much more consistently than reliability, the field would still benefit from further efforts aimed at establishing certain aspects of validity. In the section below, we briefly define and summarize different forms of validity, review relevant research, and discuss how these different forms of validity can be established in the context of neuroimaging research.

4.1. Face, content, and construct validity

When developing an emotional stress paradigm, it is important to first address face, content, and construct validity. Face validity and content validity are related in that they both look at how well a measure appears to measure what it is supposed to, but they differ based on who makes those judgments (Groth-Marnat, 2009). While face validity focuses on the perspective of the person completing the measure, content validity is concerned with expert opinions of the measure (Groth-Marnat, 2009). To a certain extent, face and content validity are both subjective in nature, but they are still an important starting point for developing a task. For example, a task purporting to measure emotional response to images/scenes could use images that are judged at face value to be emotionally relevant by participants (face validity) and experts (content validity). Construct validity is a more objective measure of validity that looks at how well a measure tests the psychological or theoretical construct it is supposed to measure (Groth-Marnat, 2009). To test construct validity, researchers can calculate correlations with measures of the same or similar constructs (e.g., for two different measures related to stress) or with different levels of responses (e.g., self-report to behavioral or physiological responses). For example, an fMRI paradigm that focuses on eliciting fearful emotions could demonstrate that participants report fear while performing the task, that individuals rating high levels of fear on this task also report high levels of fear on other behavioral or self-report tasks, and that these ratings also show a relationship with physiological responsivity (i.e., heart rate, galvanic skin response). The field has a history of identifying how activation relates to self-reported symptoms, and this is particularly true for fMRI studies of PTSD (Etkin and Wager, 2007; Hughes and Shin, 2011). However, relationships between neural activation and physiological responsivity or behavior outside of the scanner have been less of a focus. Regardless of whether an emotional stress paradigm falls within the cognitive-behavioral or symptom provocation domain (Rauch et al., 2003), it is necessary to demonstrate that these tasks are testing the psychological constructs that they are attempting to test. Note that such analyses should be conducted to confirm a priori hypotheses concerning relationships. While exploratory analyses regarding the relationship between neural activations and behavioral or clinical constructs can be useful in generating hypotheses, such relationships should be confirmed in replications. One last consideration is how factors related to validity may interact with those related to reliability. For example, using stimuli of specific relevance to a population (e.g., as in symptom provocation paradigms) or individual (e.g., personalized trauma scripts) may have a high level of validity, but may be related to lower test-retest reliability as compared to more standardized sets of stimuli (e.g., standardized emotional stimuli).

4.2. Criterion validity

Criterion validity is how well the results from a test relate to external measures or outcomes and includes both concurrent and predictive validity (Groth-Marnat, 2009). Concurrent validity is the degree to which a test relates to existing measures or “gold standards”, which in psychiatry can often be considered diagnoses or clinical symptoms. Predictive validity is the degree to which a test accurately predicts outcomes that will take place in the future (Groth-Marnat, 2009). Predictive validity is particularly relevant to the future clinical utility of emotional stress paradigms. For example, does activation during a task predict outcomes, either naturalistically or to treatment? This is very important for utility as it would shape treatment recommendations made by a clinician. An additional form of validity related to criterion validity is incremental validity, which is the ability of a measure to provide accurate results that are better than existing measures (Groth-Marnat, 2009). To establish the incremental validity of emotional stress paradigms, it would require a demonstration that these measures provide predictive or explanatory value that is better than existing measures. There have been a few studies investigating predictors of treatment response for PTSD, using emotional faces (Bryant et al., 2008; Fonzo et al., 2017), anticipation of emotional images (van Rooij et al., 2015; Aupperle et al., 2013), as well as emotional conflict and regulation tasks (Fonzo et al., 2017). Across these studies, activation within areas of the ACC/dmPFC, dorsolateral PFC, and amygdala have been highlighted as predicting response to cognitive-behavioral treatments for PTSD. However, the specific task utilized was different in all studies, the directionality of findings was variable, and the sample sizes and analytic techniques used do not yet allow for using such data to make individual-level predictions.

To progress towards clinical utility, neuroimaging research can also be aimed at identifying (a) whether paradigms developed to target a specific therapy-related process (i.e., cognitive reappraisal, fear extinction) are sensitive to those treatment effects and (b) identify novel treatment strategies to most effectively target identified neural dysfunctions. A few studies have used fMRI to investigate changes in neural response with cognitive-behavioral treatments for PTSD, utilizing emotional faces (Felmingham et al., 2007), affective Stroop (Roy et al., 2010; Thomaes et al., 2012), anticipation of emotional images (Aupperle et al., 2013), and fear conditioning, extinction and recall (Helpman et al., 2016) tasks. These studies have highlighted decreases in amygdala or insula activation as well as variable directionality of changes in rostral or dorsal ACC activation. Changes in subgenual ACC and parahippocampal gyrus were also implicated in the fear learning study, with decreased activation in these regions and increased functional coherence between PFC regions (rACC, vmPFC, sgACC) being reported during extinction recall (Helpman et al., 2016). While these studies have been conducted with varying tasks and analytic methods, it is reassuring that there is some consistency in the regions identified. However, many of these studies were conducted with relatively small samples or did not include a treatment comparison group to control for repeat testing. As we have discussed, the reliability of the tasks used is either unknown or less than sufficient and thus, may limit the generalizability of these findings. Regardless, these initial studies suggest that cognitive-behavioral and/or trauma-focused interventions may target prefrontal-insula-amygdala circuitry, offering a potential framework and hypotheses for future work.

5. Development of normative data

When a self-report or neuropsychological assessment is used clinically, the practitioner usually uses some sort of metric that compares their patient to a normative population. For example, with PTSD measures, the practitioner is likely to use normatively defined cutoffs to establish whether the person screens positive for potential PTSD diagnosis or whether they score in the mild, moderate, or severe range (Weathers et al., 2013). Neuropsychologists often establish a *z* or *t* score to represent how any one patient performed relative to a normative population (Mitrushina et al., 2005). Such a comparison provides the clinician and patient with important information on (a) whether deficits being described are more than expected given this person's demographic, and (b) the severity of these deficits. Furthermore, neuropsychological assessments often provide individuals with information on both absolute and relative strengths and weaknesses, as well as what the performance may mean in terms of daily functioning, diagnosis, treatment options, and prognosis. Similarly, clinical symptom measures are often used by clinicians to identify the targets of therapy (e.g., in identifying whether a patient's symptoms are more associated with depression versus PTSD and designing treatment accordingly).

If one is to use fMRI stress-related assessments clinically, it is difficult to imagine this being possible without normative data to which to compare individual responses. Issues with reliability or validity, as discussed in the above sections, can prevent such normative data from being meaningful. Thus, the establishment of normative data would have to come after consistent, reliable, and valid methodology has been identified (e.g., in terms of specific paradigms, contrasts analyzed, preprocessing methods, regions of interest, etc.). The establishment of normative data requires researchers to establish the distribution of responses in the general population, replicate that across samples, as well as perhaps sub-samples based on gender, age, ethnicity, and other relevant factors (e.g., socioeconomic factors; O'Connor, 1990). Such studies obviously require very large samples. Notably, there are completed or currently ongoing studies with data-sharing initiatives, which have resulted in large repositories of fMRI data (e.g., OpenfMRI, Poldrack et al., 2013, Poldrack and Gorgolewski 2017; international Study to Predict Optimized Treatment for Depression, iSPOT-D, http://www.brainresource.com/home.html, Williams et al., 2011; the NIH Adolescent Brain Cognitive Development (ABCD) Study, https://abcdstudy.org/about.html). Thus, what may have seemed impossible in the past regarding the collection of large, normative samples is now a practical reality within fMRI research. Interestingly, we are aware of only one study that was published with the explicit purpose of providing normative data in relation to functional MRI (Ball et al., 2017). The authors provide normative data for default mode network connectivity during resting state, based on the iSPOT-D study. This study, although not focused on stress-related tasks, provides a useful framework for how functional MRI research of stress responses may progress in a way that lends itself to clinical utility.

6. Roadmap for the optimal development of an fMRI paradigm

From the above review of the literature, it is obvious that there are several gaps within the field of stress-related human neuroimaging that limit its clinical utility. However, given the complications of functional neuroimaging, it can be difficult to imagine how we might fill these gaps to move forward. Here, we aspire to sketch a basic roadmap for the optimal development of emotional stress fMRI paradigms, recognizing primary obstacles for each step (see Fig. 2).

6.1. Face/content validity and internal consistency

We propose that the initial step in the development of stress-related neuroimaging paradigms is to consider *face and content validity*. This would involve obtaining input from experts in the field, as well as populations of interest (e.g., PTSD patients) during the development process (Rose et al., 2011). Depending on the purpose of the task, one would likely incorporate clinical understanding of specific symptoms or treatment-related constructs (i.e., such as was done with emotional reappraisal tasks; Buhle et al., 2014). Furthermore, one may also consider translating tasks utilized in human behavioral or animal stress-related research for use in conjunction with human neuroimaging, as has been done with fear conditioning/extinction paradigms (Hermans et al., 2006) and recently developed approach-avoidance conflict paradigms (Kirlic et al., 2017). Assessment regarding *internal consistency* of a task may begin with a focus on behavioral or self-report aspects of the task, ensuring that the stimuli used consistently and reliably invoke the expected response in individuals (e.g., in terms of self-reported ratings of valence, arousal, etc.). Trials that do not invoke the expected responses would be replaced or altogether dropped from the task. This would help to ensure that the individual trials or stimuli used in the task are consistent and robust prior to initial testing in more costly neuroimaging studies. Notably, it would then also be optimal during initial neuroimaging studies to assess for whether individual trials are consistent in their ability to invoke the expected neural response. While this is rarely done when developing fMRI-related tasks (but has been used to assess event related potentials [Olvet and Hajcak, 2009]), it could maximize the robustness (and eventual reliability) of the task as a whole. In this vein, item response theory could allow for estimating neural responses using models that incorporate individual trial discrimination parameters (Thomas et al., 2013). Inter-“item” correlations, and random sampling methods (i.e., for assessing Cronbach's Alpha) would optimally be utilized to confirm the consistency of response patterns across trials. This phase of development would be the optimal time to also develop different versions of the task if applicable, so that one could assess for consistency across trials and versions. Development of different versions of the task would lend itself well for assessment of treatment effects. Note that while the reliabilities of behavioral and neural responses are examined separately, the relationship between these responses can be explored when assessing construct validity. While one may assume that greater reliability of a behavioral response will in turn lead to greater reliability in a neural response, this may not necessarily be the case and should be explored in future studies.

6.2. Construct validity in healthy controls

As is usually the case in fMRI studies, the initial validation of a task in conjunction with neuroimaging would likely begin using healthy control populations. Such studies would not only allow for investigation of internal consistency as discussed above, but also for initial construct validation. Depending on the purpose of the task, this would either include (a) identifying that the task is associated with activation in the expected brain regions (as established, for example, by other cognitive-behavioral types of paradigms that aim to probe the similar neural systems), or (b) that task-related neural activations are associated with specific behaviors or self-report ratings relevant to the construct of interest. Although healthy control studies usually do not support investigating relationships with relevant clinical symptoms, construct validity can be assessed by examining relationships with relevant dimensional, potentially transdiagnostic psychological traits, behavioral or physiological indices collected during the fMRI paradigm, and/or other related paradigms conducted outside of the scanner. For example, as evidence of construct validity, one could view reports that vmPFC activity during fear extinction related to an extinction retention index based on SCRs (Milad et al., 2007). We propose that it may be *optimal* for tasks to have concurrent behavioral or physiological indices, which would not only be used for validation, but would also continue to be

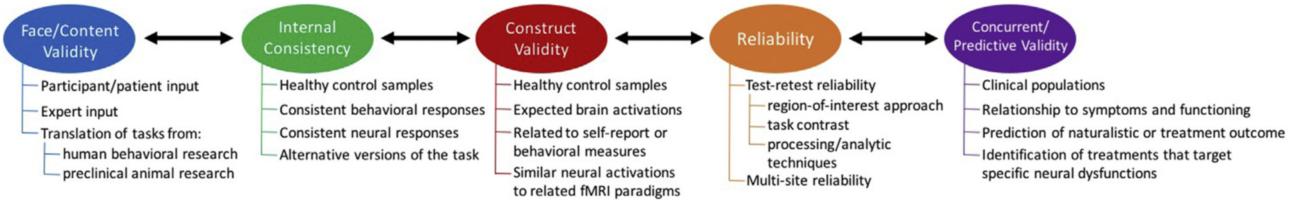
Face/Content Validity
Participant/patient input
Expert input
Translation of tasks from:
human behavioral research
preclinical animal research
Internal Consistency
Healthy control samples
Consistent behavioral responses
Consistent neural responses
Alternative versions of the task
Construct Validity
Healthy control samples
Expected brain activations
Related to self-report or behavioral measures
Similar neural activations to related fMRI paradigms
Reliability
Test-retest reliability
region-of-interest approach
task contrast
processing/analytic techniques
Multi-site reliability
Concurrent/ Predictive Validity
Clinical populations
Relationship to symptoms and functioning
Prediction of naturalistic or treatment outcome
Identification of treatments that target specific neural dysfunctions

Fig. 2. Clinical Utility Roadmap for fMRI Emotional Stress Paradigms.
This diagram outlines a “roadmap” for researchers to use as a guide for designing methodologically sound emotional stress paradigms for fMRI studies. Note that between these steps are bidirectional arrows, recognizing the possibility and likelihood that these steps may not be completed in a linear fashion and that steps may need to be repeated during the development process.

used alongside the task during further development. This would increase the likelihood of identifying cheaper analogues to neural activation patterns, which can subsequently be disseminated more easily than fMRI to clinical contexts. Furthermore, this information would be crucial during the construct validation stage to make any necessary task modifications. Thus, if the task is not serving its initial purpose of probing specific circuits, or tapping into a specific construct, the task could be modified to do so more effectively.

6.3. Test-retest reliability

After developing tasks that have face validity, internal consistency, and initial construct validity in healthy control samples, the next optimal step would be to conduct studies to examine the test-retest and perhaps even multi-site reliability of the task. This step is time and resource intensive and is therefore often ignored within fMRI research. However, if a task does not consistently produce similar patterns of neural activation, clinical interpretations on an individual-level cannot be made. In addition, reliable tasks are likely to be more robust when used in the context of prediction or treatment response (Lachin, 2004). As discussed in this review, there have been several test-retest fMRI studies (and a few multi-site reliability studies) focused on emotional face tasks. However, there has been a surprising lack of such studies with other stress- or emotion-related tasks. Even with emotional faces tasks, the ICCs reported by test-retest reliability measures were highly variable (see Table 1). We propose that such reliability studies, rather than simply aiming to test whether or not the *commonly* used metrics of the task are reliable, focus on identifying the specific metrics that do in fact demonstrate adequate reliability. FMRI response is not a singular measure, but rather a conglomerate of measures from various groupings of voxels or regions of interest. Moreover, most tasks examine a range of potential contrasts of interest. Therefore, it is quite possible to identify the specific indices that evidence most reliability, as has been done with functionally-defined amygdala habituation to faces-shapes contrast for emotional faces tasks by Plichta et al. (2014) and with graph theory by Cao et al. (2014). The use of factor analytic or other data reduction techniques (e.g., independent component analysis; Congdon et al., 2010) could be used to identify latent factors or components representing brain responses to emotional stress tasks (as has been done with resting state in conjunction with intervention research; Sripada et al., 2013). Such factors may be more stable than individual measures, as a change between testing sessions in a few voxels or even one brain region is going to have a limited impact on the overall factor. Computational approaches (such as Bayesian approaches) have the potential for modeling complex behavioral and neural responses in ways that further elucidates our understanding of psychiatric symptomatology (Huys et al., 2016), and which may also prove to be more stable over time than traditional measures.

It would be further optimal to conduct multiple test-retest studies, with different time frames in between (i.e., a few days, a few weeks, a few months). However, if the task is going to be used to assess for treatment effects or to determine whether a treatment targets a specific neural network, test-retest studies should be designed to match the typical length of treatment. Another issue with test-retest reliability is whether there are environmental variables that need to be controlled for in order to increase reliability. In the cardiovascular field for example, it is often specified what environmental variables should be controlled for (e.g., caffeine/tobacco/food intake, menstrual phase, room temperature, etc.; Harris et al., 2010). In fMRI research related to stress, the control of these types of variables, as well as data processing steps, may help to increase test-retest reliability. However, practical considerations for implementing environmental control should not be ignored (ease of scheduling; burden on participants). It is ultimately important to establish measures that are reliable in the settings in which they would be administered.

6.4. Normative data

Once a task is developed and specific indices have been identified that demonstrate evidence of construct validity and adequate test-retest reliability, the collection of normative data would be an important next step. As mentioned previously, fMRI researchers have not previously focused on the publication of normative data. While we suggest this as a step that would optimally be conducted prior to research with clinical populations, we recognize this is often not going to be the case. Normative data requires large samples sizes and researchers likely will hesitate to invest resources towards this until they have established that the task indices are clinically meaningful. Regardless of when such normative data is established, it is a necessary and powerful step towards using the indices to make meaningful interpretations at the individual level.

6.5. Criterion validity

The last step in development of clinically useful stress-related neuroimaging paradigms would be to establish criterion validity in patient samples. Cross-sectional research has offered insights into potential neural systems relating to psychiatric symptomatology and are important steps in establishing concurrent validity of the task indices. However, it is difficult with cross-sectional research to establish predictive validity, that is, whether findings represent deficits or compensatory responses that are adaptive for long-term functioning. These studies also cannot determine whether or not commonly used treatment modalities already target the identified neural dysfunctions effectively. Thus, longitudinal research addressing predictive validity is perhaps one of the most important steps in moving the neuroimaging field toward meaningful clinical utility. As noted above, there have been a few studies conducted to identify neural predictors and changes related to cognitive-behavioral therapy for PTSD. However, due to small sample sizes, limits in statistical approaches, and inconsistencies in findings across studies, there remains a lack of obvious translation of findings to individual patients in the clinical realm. In predictive research, the use of machine learning techniques with separate participant samples for training and replication of the predictive model will represent substantial progress in this regard (Ball et al., 2014; Mansson et al., 2015). Note that while these predictive models are providing evidence of predictive validity, this replication across samples could also be considered a form of reliability. However, eventually, prospective research will likely be needed to examine the effectiveness of such a predictive test for enhancing outcomes via treatment selection (Mandrekar and Sargent, 2009). It is also important for researchers to establish the accuracy and specificity for any given predictor and establish that it is more meaningful than simply measuring symptom severity or diagnosis, which would provide evidence of incremental validity. This too, is a step in which identification of behavioral, physiological, or self-report variables that can be used to estimate the neural response profiles would be beneficial for making such findings practical and able to be disseminated in the clinical setting.

Given the infinite number of different treatment modalities, the goal of identifying neural predictors of treatment response or targets of various treatment strategies can seem overwhelming. However, it is important to recognize that we are currently in a place where clinicians have very little information on which to guide their treatment decisions. Thus, identifying a measure that can robustly and reliably predict how well an individual person would likely respond to even just one specific intervention (e.g., cognitive restructuring techniques), would in and of itself be useful. Identifying a brief intervention that is known to target specific neural dysfunctions when observed in a patient (e.g., to enhance recruitment of dorsolateral PFC regions during emotion regulation), would offer a meaningful place for clinicians to start.

6.6. Roadmap summary

None of the paradigms used in stress-related research have been developed in such a sequential and thorough way. Emotional faces tasks have undergone the most scrutiny and are the most widely used in reliability studies, yet the measures often relied upon in the majority of studies are not necessarily the measures found to be most reliable (Cao et al., 2014; Plichta et al., 2012, 2014). Measurements focusing on habituation or graph theory have been mentioned preciously, but other potential methods that should be explored as potentially more stable measurements include multivariate pattern analysis, which examines how patterns of voxels directly relate to the psychological construct of interest (e.g., pain intensity; Woo et al., 2017a), or machine-learning measures including support vector machine (SVM) datasets that can be transferred across populations (Woo et al., 2017b). However, it is also important to consider how these alternative methods (e.g., SVM) could add undue complication that make findings less interpretable for clinicians attempting to utilize these measures. An additional consideration for more traditional methodological approaches relates to the size of an ROI used in any given study and how that affects reliability. Researchers should consider the size of the functional activation they are attempting to capture and use appropriate ROIs that will accurately capture this. There is an obvious need within the field for more effort, energy, and resources to be spent on establishing reliable measures that provide clinically meaningful information. Importantly, this does not necessarily mean that researchers should start from scratch in terms of paradigm development. On the contrary, researchers can leverage data already collected as part of large, multi-site trials with data-sharing initiatives and can conduct rigorous test-retest reliability studies to identify the most stable measures associated with currently-used paradigms. The proposed "roadmap" is therefore often likely to be a winding road in which researchers circle-back to fill in methodological gaps.

7. Conclusion

Stress-related neuroimaging tasks have perhaps the most potential for providing information about individual differences in neural processing that may be clinically useful for PTSD and other mental health disorders. However, this potential clinical utility is limited by gaps in methodological development. This review paper highlights research that has been conducted with stress-related human neuroimaging paradigms to establish basic reliability and validity. The emotional faces paradigm, which involves neural circuits of interest to stress-related mental health populations (i.e., amgydala), has received the most attention in this regard. Studies suggest that using specific contrasts and regions of interest with this task may increase test-retest and multi-site reliability. Other tasks, such as symptom provocation and fear learning tasks, have not been well scrutinized in terms of the reliability of neural response patterns. We outline a roadmap towards optimizing the potential clinical utility of neuroimaging tasks, which involves consideration of face/content and construct validity, test-retest and multi-site reliability, development of normative datasets, and concurrent and predictive validity (Fig. 2). By addressing these methodological gaps, we will likely have a powerful effect on improving the replicability of findings and optimize our chances for improving real-world clinical outcomes.

Acknowledgements

Robin Aupperle, Ph.D. received support from the National Institute of Mental Health of the National Institutes of Health under Award Number K23MH108707.

References

Aupperle, R.L., Allard, C.B., Simmons, A.N., Flagan, T., Thorp, S.R., et al., 2013. Neural responses during emotional processing before and after cognitive trauma therapy for battered women. Psychiatr. Res. 214 (1), 48–55.

Ball, T.M., Goldstein-Piekarski, A.N., Gatt, J.M., Williams, L.M., 2017. Quantifying person-level brain network functioning to facilitate clinical translation. Transl. Psychiatry 7 (10), e1248.

Ball, T.M., Stein, M.B., Ramsawh, H.J., Campbell-Sills, L., Paulus, M.P., 2014. Single-subject anxiety treatment outcome prediction using functional neuroimaging. Neuropsychopharmacology 39 (5), 1254–1261.

Bartko, J.J., 1976. On various intraclass correlation reliability coefficients. Psychol. Bull. 83 (5), 762.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155.

Blair, K., Shaywitz, J., Smith, B.W., Rhodes, R., Geraci, M., et al., 2008. Response to emotional expressions in generalized social phobia and generalized anxiety disorder: evidence for separate disorders. Am. J. Psychiatr. 165 (9), 1193–1202.

Boehme, S., Mohr, A., Becker, M.P., Miltner, W.H., Straube, T., 2014. Area-dependent time courses of brain activation during video-induced symptom provocation in social anxiety disorder. Biol. Mood Anxiety Disord. 4, 6.

Breiter, H.C., Etcoff, N.L., Whalen, P.J., Kennedy, W.A., Rauch, S.L., 1996. Response and habituation of the human amygdala during visual processing of facial expression. Neuron 17 (5), 875–887.

Brennan, R.L., 2001. Generalizability Theory. Springer, New York.

Britton, J.C., Phan, K.L., Taylor, S.F., Fig, L.M., Liberzon, I., 2005. Corticolimbic blood flow in posttraumatic stress disorder during script-driven imagery. Biol. Psychiatr. 57 (8), 832–840.

Brown, G.G., Mathalon, D.H., Stern, H., Ford, J., Mueller, B., et al., 2011. Multisite reliability of cognitive BOLD data. NeuroImage 54 (3), 2163–2175.

Bryant, R.A., Felmingham, K., Kemp, A., Das, P., Hughes, G., et al., 2008. Amygdala and ventral anterior cingulate activation predicts treatment response to cognitive behavior therapy for post-traumatic stress disorder. Psychol. Med. 38 (4), 555–561.

Buhle, J.T., Silvers, J.A., Wager, T.D., Lopez, R., Onyemekwu, C., et al., 2014. Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. Cereb. Cortex 24 (11), 2981–2990.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., et al., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.

Cannon, T.D., Cao, H., Mathalon, D.H., Gee, D.G., 2018. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study: clarification and implications for statistical power. Hum. Brain Mapp. 39 (1), 599–601.

Cao, H., Plichta, M.M., Schafer, A., Haddad, L., Grimm, O., et al., 2014. Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. NeuroImage 84, 888–900.

Carter, C.S., Heckers, S., Nichols, T., Pine, D.S., Strother, S., 2008. Optimizing the design and analysis of clinical functional magnetic resonance imaging research studies. Biol. Psychiatr. 64, 842–849.

Congdon, E., Mumford, J.A., Cohen, J.R., Galvan, A., Aron, A.R., et al., 2010. Engagement of large-scale networks is related to individual differences in inhibitory control. NeuroImage 53 (2), 653–663.

Davis, M., Whalen, P.J., 2001. The amygdala: vigilance and emotion. Mol. Psychiatr. 6 (1), 13–34.

Etkin, A., Wager, T.D., 2007. Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. Am. J. Psychiatr. 164 (10), 1476–1488.

Fanselow, M.S., Ponnusamy, R., 2008. Chapter 2.2 the use of conditioning tasks to model fear and anxiety. Handb. Behav. Neurobiol. 17, 29–48.

Felmingham, K., Kemp, A., Williams, L., Das, P., Hughes, G., et al., 2007. Changes in anterior cingulate and amygdala after cognitive behavior therapy of posttraumatic stress disorder. Psychol. Sci. 18 (2), 127–129.

Fischer, H., Wright, C.I., Whalen, P.J., McInerney, S.C., Shin, L.M., Rauch, S.L., 2003. Brain habituation during repeated exposure to fearful and neutral faces: a functional MRI study. Brain Res. Bull. 59 (5), 387–392.

Fleiss, J.L., 1986. Analysis of data from multiclinic trials. Control Clin Trials 7, 267–275.

Fonzo, G.A., Goodkind, M.S., Oathes, D.J., Zaiko, Y.V., Harvey, M., et al., 2017. PTSD psychotherapy outcome predicted by brain activation during emotional reactivity and PTSD psychotherapy outcome predicted by brain activation during emotional reactivity and regulation. Am. J. Psychiatr. 174 (12), 1163–1174.

Fredrikson, M., Annas, P., Georgiades, A., Hursti, T., Tersman, Z., 1993. Internal consistency and temporal stability of classically condition skin conductance responses. Biol. Psychol. 35, 153–163.

Fullana, M.A., Harrison, B.J., Soriano-Mas, C., Vervliet, B., Cardoner, N., et al., 2015. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. Mol. Psychiatr. 21 (4), 500–508.

Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., et al., 2009. Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. J. Psychiatry Neurosci. 34 (6), 418–432.

Gee, D.G., McEwen, S.C., Forsyth, J.K., Haut, K.M., Bearden, C.E., et al., 2015. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study. Human Brain Mapp. 36 (7), 2558–2579.

Groth-Marnat, G., 2009. Handbook of Psychological Assessment. John Wiley & Sons, Hoboken, NJ.

Haaker, J., Golkar, A., Selbing, I., Olsson, A., 2017. Assessment of social transmission of threats in humans using observational fear conditioning. Nat. Protoc. 12 (7), 1378–1386.

Hariri, A.R., Tessitore, A., Mattay, V.S., Fera, F., Weinberger, D.R., 2002. The amygdala response to emotional stimuli: a comparison of faces and scenes. NeuroImage 17, 317–323.

Harris, R.A., Nishiyama, S.K., Wray, D.W., Richardson, R.S., 2010. Ultrasound assessment of flow-mediated dilation. Hypertension 55 (5), 1075–1085.

Helpman, L., Marin, M.F., Papini, S., Zhu, X., Sullivan, G.M., et al., 2016. Neural changes in extinction recall following prolonged exposure treatment for PTSD: a longitudinal fMRI study. NeuroImage Clin. 12, 715–723.

Hendler, T., Rotshtein, P., Yeshurun, Y., Weizmann, T., Kahn, I., et al., 2003. Sensing the invisible: differential sensitivity of visual cortex and amygdala to traumatic context. NeuroImage 19, 587–600.

Hermans, D., Craske, M.G., Mineka, S., Lovibond, P.F., 2006. Extinction in human fear conditioning. Biol. Psychiatr. 60 (4), 361–368.

Hopper, J.W., Frewen, P.A., van der Kolk, B.A., Lanius, R.A., 2007. Neural correlates of reexperiencing, avoidance, and dissociation in PTSD: symptom dimensions and emotion dysregulation in responses to script-driven trauma imagery. J. Trauma Stress 20 (5), 713–725.

Hughes, K.C., Shin, L.M., 2011. Functional neuroimaging studies of post-traumatic stress disorder. Expert Rev. Neurother. 11 (2), 275–285.

Huys, Q.J.M., Maia, T.V., Paulus, M.P., 2016. Computational psychiatry: from mechanistic insights to the development of new treatments. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 1 (5), 382–385.

Hygge, S., Ohman, A., 1978. Modeling processes in the acquisition of fears: vicarious electrodermal conditioning to fear-relevant stimuli. J. Pers. Soc. Psychol. 36 (3), 271–279.

Infantolino, Z.P., Luking, K.R., Sauder, C.L., Curtin, J.J., Hajcak, G., 2018. Robust is not necessarily: from within-subjects fMRI contrasts to between-subjects comparisons. NeuroImage 173, 146–152.

Johnstone, T., Somerville, L.H., Alexander, A.L., Oakes, T.R., Davidson, R.J., et al., 2005. Stability of amygdala BOLD response to fearful faces over multiple scan sessions. NeuroImage 25, 1112–1123.

Kirlic, N., Young, J., Aupperle, R.L., 2017. Animal to human translational paradigms relevant for approach avoidance conflict decision making. Behav. Res. Ther. 96, 14–29.

Lachin, J.M., 2004. The role of measurement reliability in clinical trials. Clin. Trials 1 (6), 553–566.

Lahey, M.A., Downey, R.G., Saal, F.E., 1983. Intraclass correlations: there's more than meets the eye. Psychol. Bull. 93 (3), 586.

Lang, P.J., Bradley, M.M., Cuthbert, B.N., 2008. International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. Technical Report A-8. University of Florida, Gainesville, FL.

Lanius, R.A., Williamson, P.C., Densmore, M., Boksman, K., Madhulika, G.A., et al., 2001. Neural correlates of traumatic memories in posttraumatic stress disorder: a functional MRI investigation. Am. J. Psychiatr. 158, 1920–1922.

Lieberman, M.D., Eisenberger, N.I., Crockett, M.J., Tom, S.M., Pfeifer, J.H., Way, B.M., 2007. Affect labeling disrupts amygdala activity in response to affective stimuli. Psychol. Sci. 18 (5), 421–428.

Liberzon, I., Abelson, J.L., 2016. Context processing and the neurobiology of post-traumatic stress disorder. Neuron 92 (1), 14–30.

Lipp, I., Murphy, K., Wise, R.G., Caseras, X., 2014. Understanding the contribution of neural and physiological signal variation to the low repeatability of emotion-induced BOLD responses. NeuroImage 86, 335–342.

Lonsdorf, T.B., Menz, M.M., Andreatta, M., Fullana, M.A., Golkar, A., 2017. Don't fear 'fear conditioning': methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. Neurosci. Biobehav. Rev. 77, 247–285.

Mandrekar, S.J., Sargent, D.J., 2009. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J. Clin. Oncol. 27 (24), 4027–4034.

Mansson, K.N.T., Frick, A., Boraxbeek, C.J., Marquand, A.F., Williams, S.C., et al., 2015. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. Transl. Psychiatry 5 (3), e530.

Manuck, S.B., Brown, S.M., Forbes, E.E., Hariri, A.R., 2007. Temporal stability of individual differences in amygdala reactivity. Am. J. Psychiatr. 164 (10), 1613–1614.

McClean, C.P., Foa, E.B., 2013. Dissemination and implementation of prolonged exposure therapy for posttraumatic stress disorder. J. Anxiety Disord. 27, 788–792.

Mechias, M.L., Etkin, A., Kalisch, R., 2010. A meta-analysis of instructed fear studies: implications for conscious appraisal of threat. NeuroImage 49, 1760–1768.

Milad, M.R., Rosenbaum, B.L., Simon, N.M., 2014. Neuroscience of fear extinction: implications for assessment and treatment of fear-based and anxiety related disorders. Behav. Res. Ther. 62, 17–23.

Milad, M.R., Wright, C.I., Orr, S.P., Pitman, R.K., Quirk, G.J., Rauch, S.L., 2007. Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. Biol. Psychiatr. 62 (5), 446–454.

Mitrushina, M., Boone, K.B., Razani, J., D'Elia, L.F., 2005. Handbook of Normative Data for Neuropsychological Assessment. Oxford University Press, New York, NY.

Nitschke, J.B., Sarinopoulos, I., Mackiewicz, K.L., Schaefer, H.S., Davidson, R.J., 2006. Functional neuroanatomy of aversion and its anticipation. NeuroImage 29, 106–116.

Nord, C.L., Gray, A., Charpentier, C.J., Robinson, O.J., Roiser, J.P., 2017. Unreliability of putatitve fMRI biomarkers during emotional face processing. NeuroImage 156, 119–127.

O'Connor, P.J., 1990. Normative data: their definition, interpretation, and importance for primary care physicians. Fam. Med. 22 (4), 307–311.

Ochsner, K.N., Bunge, S.A., Gross, J.J., Gabrieli, J.D.W., 2002. Rethinking feelings: an FMRI study of the cognitive regulation of emotion. J. Cogn. Neurosci. 14 (8), 1215–1229.

Olsson, A., Phelps, E.A., 2007. Social learning of fear. Nat. Neurosci. 10 (9), 1095–1102.

Olvet, D.M., Hajcak, G., 2009. The stability of error-related brain activity with increasing trials. Psychophysiology 46 (5), 957–961.

Paulus, M.P., 2015. Pragmatism instead of mechanism A call for impactful biological psychiatry. JAMA Psychiatry 72 (7), 631–632.

Pavlov, I.P., 1927. Conditioned Reflexes: an Investigation of the Physiological Activity of the Cerebral Cortex. Oxford Univ. Press, Oxford, England.

Phan, K.L., Wager, T.D., Taylor, S.F., Liberzon, I., 2002. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. NeuroImage 16, 331–348.

Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., et al., 2012. Test-retest reliability of evoked BOLD signals from a cognitive-emotiove fMRI test battery. NeuroImage 60, 1746–1758.

Plichta, M.M., Grimm, O., Morgen, K., Mier, D., Sauer, C., et al., 2014. Amygdala habituation: a reliable fMRI phenotype. NeuroImage 103, 383–390.

Poldrack, R.A., Barch, D.M., Mitchell, J.P., Wager, T.D., Wagner, A.D., et al., 2013. Toward open sharing of task-based fMRI data: the OpenfMRI project. Front Neuroinform. 7, 12.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., et al., 2016. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18, 115–126.

Poldrack, R.A., Gorgolewski, K.J., 2017. OpenfMRI: open sharing of task fMRI data. NeuroImage 144 (Pt B), 259–261.

Posamentier, M.T., Abdi, H., 2003. Processing faces and facial expressions. Neuropsychol. Rev. 13 (3), 113–143.

Rabinak, C.A., MacNamara, A., Kennedy, A.E., Angstadt, M., Stein, M.B., 2014. Focal and aberrant prefrontal engagement during emotion regulation in veterans with posttraumatic stress disorder. Depress. Anxiety 31 (10), 851–861.

Rauch, S.L., Shin, L.M., Wright, C.I., 2003. Neuroimaging studies of amygdala function in anxiety disorders. Ann. N. Y. Acad. Sci. 985, 389–410.

Riboni, F.V., Belzung, C., 2017. Stress and psychiatric disorders: from categorical to dimensional approaches. Curr. Opin. Behav. Sci. 14, 72–77.

Robinson, O.J., Charney, D.R., Overstreet, C., Vytal, K., Grillon, C., 2012. The adaptive threat bias in anxiety: amygdala–dorsomedial prefrontal cortex coupling and aversive amplification. NeuroImage 60, 523–529.

Rose, D., Evans, J., Sweeney, A., Wykes, T., 2011. A model for developing outcome measures from the perspectives of mental health service users. Int. Rev. Psychiatry 23 (1), 41–46.

Roy, M.J., Francis, J., Friedlander, J., Banks-Williams, L., Lande, R.G., 2010. Improvement in cerebral function with treatment of posttraumatic stress disorder. Ann. N.Y. Acad. Sci. 1208, 142–149.

Sartory, G., Cwik, J., Knuppertz, H., Schurholt, B., Lebens, M., et al., 2013. PLoS One 8 (3) e58150.

Sauder, C.L., Hajcak, G., Angstadt, M., Phan, K.L., 2013. Test-retest reliability of amygdala response to emotional faces. Psychophysiology 50 (11), 1146–1156.

Schacher, M., Haemmerle, B., Woermann, F.G., Okujava, M., Huber, D., et al., 2006. Amygdala fMRI lateralizes temporal lobe epilepsy. Neurology 66 (1), 81–87.

Schienle, A., Schafer, A., Hermann, A., Rohrmann, S., Vaitl, D., 2007. Symptom provocation and reduction in patients suffering from spider phobia: an fMRI study on exposure therapy. Eur. Arch. Psychiatry Clin. Neurosci. 257 (8), 486–493.

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.

Schunck, T., Erb, G., Mathis, A., Jacob, N., Gilles, C., et al., 2008. Test-retest reliability of a functional MRI anticipatory anxiety paradigm in healthy volunteers. J. Magn. Reson. Imaging 27, 459–468.

Shafran, R., Clark, D.M., Fairburn, C.G., Arntz, A., Barlow, D.H., 2009. Mind the gap: improving the dissemination of CBT. Behav. Res. Ther. 47, 902–909.

Shavelson, R.J., Webb, N.M., 1991. Generalizability Theory: a Primer. SAGE, Thousand Oaks, CA.

Shin, L.M., Liberzon, I., 2010. The neurocircuitry of fear, stress, and anxiety disorders. Neuropsychopharmacology 35, 169–191.

Shrout, P.E., Fleiss, J.E., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

Simmons, A., Strigo, I., Matthews, S.C., Paulus, M.P., Stein, M.B., 2006. Anticipation of aversive stimuli is associated with increased insula activation in anxiety-prone subjects. Biol. Psychiatr. 60 (4), 402–409.

Southam-Gerow, M.A., Rodriguez, A., Chorpita, B.F., Daleiden, E.L., 2012. Dissemination and implementation of evidence based treatments for youth: challenges and recommendations. Prof. Psychol. Res. Pr. 43 (5), 527–534.

Sripada, C.S., Kessler, D., Welsh, R., Angstadt, M., Liberzon, I., et al., 2013. Distributed effects of methylphenidate on the network structure of the resting brain: a connectomic pattern classification analysis. NeuroImage 81, 213–221.

Stark, R., Schienle, A., Walter, B., Kirsch, P., Blecker, C., et al., 2004. Hemodynamic effects of negative emotional pictures – a test-retest analysis. Neuropsychobiology 50, 108–118.

Thomaes, K., Dorrepaal, E., Draijer, N., de Ruiter, M.B., Elzinga, B.M., et al., 2012. Treatment effects on insular and anterior cingulate cortex activation during classic and emotional Stroop interference in child abuse-related complex post-traumatic stress disorder. Psychol. Med. 42 (11), 2337–2349.

Thomas, M.L., Brown, G.G., Thompson, W.K., Voyvodic, J., Greve, D.N., et al., 2013. An application of item response theory to fMRI data: prospects and pitfalls. Psychiatr. Res. 212 (3), 167–174.

Thomas, M.L., Patt, V.M., Bismark, A., Sprock, J., Tarasenko, M., et al., 2017. Evidence of systematic attenuation in the measurement of cognitive deficits in schizophrenia. J. Abnorm. Psychol. 126 (3), 312–324.

Torrents-Rodas, D., Fullana, M.A., Bonillo, A., Andion, O., Molinuevo, B., et al., 2014. Testing the temporal stability of individual differences in the acquisition and generalization of fear. Psychophysiology 51, 697–705.

Tottenham, N., Tanaka, J.W., Leon, A.C., McCarry, T., Nurse, M., et al., 2009. The NimStim set of facial expressions: judgments from untrained research participants. Psychiatr. Res. 168, 242–249.

van den Bulk, B.G., Koolschijn, P.C., Meens, P.H., van Lang, N.D., van der Wee, N.J., et al., 2013. How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. Developmental Cognitive Neuroscience 4, 65–76.

VanElzakker, M.B., Dahlgren, M.K., Davis, F.C., Dubois, S., Shin, L.M., 2014. From Pavlov to PTSD: the extinction of conditioned fear in rodents, humans, and anxiety disorders. Neurobiol. Learn. Mem. 113, 3–18.

van Rooij, S.J., Geuze, E., Kennis, M., Rademaker, A.R., Vink, M., 2015. Neural correlates of inhibition and contextual cue processing related to treatment response in PTSD. Neuropsychopharmacology 40 (3), 667–675.

Weathers, F.W., Blake, D.D., Schnurr, P.P., Kaloupek, D.G., Marx, B.P., Keane, T.M., 2013. The Clinician-administered PTSD Scale for DSM-5 (CAPS-5). Interview available from the National Center for PTSD at. www.ptsd.va.gov.

Webb, N.M., Shavelson, R.J., 2005. Generalizability theory: overview. In: Everitt, B.S., Howell, D.C. (Eds.), Encyclopedia of Statistics in Behavioral Science. Wiley, pp. 717–719.

Weir, J.P., 2005. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Condit Res. 19 (1), 231–240.

Whalen, P.J., Rauch, S.L., Etcoff, N.L., McInerney, S.C., Lee, M.B., Jenike, M.A., 1998. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. J. Neurosci. 18 (1), 411–418.

Whalen, P.J., Shin, L.M., McInerney, S.C., Fischer, H., Wright, C.I., Rauch, S.L., 2001. A functional MRI study of human amygdala responses to facial expressions of fear versus anger. Emotion 1 (1), 70–83.

Williams, L.M., Rush, A.J., Koslow, S.H., Wisniewski, S.R., Cooper, N., et al., 2011. International Study to Predict Optimized Treatment for Depression (iSPOT-D), a randomized clinical trial: rationale and protocol. Trials 12 (4), 1–17.

Woo, C.W., Schmidt, L., Krishnan, A., Jepma, M., Roy, M., et al., 2017a. Quantifying cerebral contributions to pain beyond nociception. Nat. Commun. 8, 14211.

Woo, C.W., Chang, L.J., Lindquist, M.A., Wager, T.D., 2017b. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. 20 (3), 365–377.

Wright, C.I., Fischer, H., Whalen, P.J., McInerney, S.C., Shin, L.M., Rauch, S.L., 2001. Differential prefrontal cortex and amygdala habituation to repeatedly presented emotional stimuli. Neuroreport 12, 379–383.

Yen, M., Lo, L.H., 2002. Examining test-retest reliability: an intra-class correlation approach. Nurs. Res. 51 (1), 59–62.

Zeidan, M.A., Lebron-Milad, K., Thompson-Hollands, J., Im, J.J., Dougherty, D.D., et al., 2012. Test-retest reliability during fear acquisition and fear extinction in humans. CNS Neurosci. Ther. 18, 313–317.

Zorn, J.V., Schur, R.R., Boks, M.P., Kahn, R.S., Joels, M., Vinkers, C.H., 2017. Cortisol stress reactivity across psychiatric disorders: a systematic review and meta-analysis. Psychoneuroendocrinology 77 (2017), 25–36.