**Distribution Agreement**

Signature:

_________________________________ ______________

Andrew Petrich Teer Date

Investigating and Developing a Novel Implicit Measurement of Self-Esteem

By

Andrew P. Teer
Master of Arts

Clinical Psychology

________________________________________
Michael T. Treadway, Ph.D.
Advisor

________________________________________
Joseph R. Manns, Ph.D.
Committee Member

________________________________________
W. Edward Craighead, Ph.D.
Committee Member

Accepted:

________________________________________
Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_________________________
Date

Investigating and Developing a Novel Implicit Measurement of Self-Esteem

By

Andrew Petrich Teer  
B.A. Yale University, 2010

Advisor: Michael T. Treadway, Ph.D.

An abstract of a thesis submitted to the faculty of the  
James T. Laney School of Graduate Studies of Emory University  
in partial fulfillment of the requirements for the degree of  
Master of Arts in Psychology  
2016

# Abstract

Investigating and Developing a Novel Implicit Measurement of Self-Esteem
By Andrew Petrich Teer

Major Depressive Disorder (MDD) is a major public health risk and its frontline treatments are only effective in half of patients or less. Depression, like most psychological disorders, is characterized by symptoms that can only be measured by patient self-report as opposed to observable, objective signs. Several symptoms critical to the construct of MDD are: suicidal thoughts, depressed mood and feelings of worthlessness. One obstacle in the development of effective treatments of depression arises from reliance on self-report measurements of these symptoms. Both demand characteristics and the inability for patients to aggregate information accurately for these symptoms may make self-reports not wholly meaningful. In particular, growing evidence suggests that self-perception and self-esteem may be difficult constructs to assess using common self-report measures alone. Implicit measurements have been developed in order to bypass some of the heuristics and biases involved in retrospective self-report, but, overall, have only been mildly successful in doing so. Tasks like the Implicit Association Task (IAT) have helped to address these limitations, but the neural circuitry involved in IAT tasks is unknown. In contrast, the neural mechanisms of fear conditioning are very well characterized. The three studies herein attempt to take advantage of well-characterized conditioning paradigms to test the feasibility of fear conditioning and extinction to self-related imagery as an objective biomarker of self-esteem. Participants completed self-report measures related to self-esteem, self-compassion, depression and anxiety. They also had their photo taken at the beginning of the study. Three photographs served as conditioned stimuli (CS), one of which was the photograph of the participant (CS-Self). The CS-Self and another photograph (CS+) were always paired with the unconditioned stimulus (UCS). A third (CS–) was never paired with the UCS. The UCS was a loud unpleasant noise. Pupil dilation served as the unconditioned response (UCR), which is a well-validated metric for autonomic arousal. The results from Study 1 indicated that participants responded the same to the CS+ and the CS–, which indicated that they did not acquire a conditioned response. The results from Studies 2 and 3 demonstrate the feasibility of using self-related imagery as the conditioned stimulus in a fear-conditioning paradigm when the stimuli are presented on a luminance-matched gray background and are not preceded by a fixation cross. Furthermore, the results of Study 3 may indicate that self-compassion could affect the fear acquisition learning. These preliminary findings support the potential for this paradigm to be used as an objective measure of self-esteem.

# Investigating and Developing a Novel Implicit Measurement of Self-Esteem

By

Andrew Petrich Teer
B.A. Yale University, 2010

Advisor: Michael T. Treadway, Ph.D.

A thesis submitted to the faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Arts in Psychology
2016

## Table of Contents

## Background and Introduction

Major Depressive Disorder (MDD) is a serious public health risk and is linked to significant impairment and disability. Worldwide, MDD is the second leading cause of disability (Ferrari et al, 2013). In a 2010 morbidity and mortality report, the Centers for Disease Control estimated that 9.1% of US citizens meet criteria for any depressive disorder, with between 4 and 6% meeting for MDD (Centers for Disease Control, 2010; Kessler *et al.*, 2009). Moreover, individuals with MDD have increased risks of both attempting and completing suicide (Mann & Currier, 2008). Furthermore, front-line treatments like Cognitive Behavior Therapy (CBT) and Interpersonal Therapy (IPT) seem to be efficacious only in one third to one half of the population (Cuijpers, *et al.*, 2008). A study comparing the efficacy of medication and cognitive therapy (CT) in depressed patients demonstrated relapse rates of approximately thirty percent in patients who received CT for sixteen weeks and seventy percent of patients who were treated with medication for sixteen weeks (Hollon *et al.*, 2005). The Diagnostic and Statistical Manual, Fourth Edition (DSM-IV) indicated that relapse rates hover around fifty percent for those suffering their first episode while approaching eighty percent for those with multiple episodes (American Psychiatric Association, 2000). The empirical support for psychopharmacological intervention is not any more promising: NIMH data suggest that selective serotonin reuptake inhibitors (SSRIs) are aggregately about as effective as placebo in the treatment depression (Insel, 2011). This ineffectiveness could be due to the fact that antidepressants are one of the most-prescribed classes of drugs, with approximately 10% of United States adults reporting having taken an antidepressant in the last month (National Center for Health Statistics, 2013).

In psychology, symptom assessment is largely completed by means of self-report measurements. The medical field's distinction between signs (objective evidence like high blood sugar, which a patient cannot always detect) and symptoms (like a headache, which can only be self-reported) is a useful dichotomy to carry into mental health diagnoses.  Such symptoms are not readily accessible through other means of assessment. The traits that characterize MDD are largely symptoms. Although signs like psychomotor retardation and hypersomnia can be observed objectively, critical symptoms like depressed mood, suicidal ideation, and feelings of worthlessness or excessive guilt can only be reported by the patient.

Self-esteem plays a critical role in the development of depression (Sowislo & Orth, 2013), as well as other disorders that are commonly co-morbid with depression such as social anxiety and personality disorders (American Psychiatric Association, 2000). Poor self-esteem is both a symptom of depression and a factor in the onset and maintenance of depression (Beck, 1987). Those with low self-esteem tend to ruminate more about negative aspects of their life, and rumination has been correlated with increased risk of depression (Nolen-Hoeksema, 2000). Interpersonal theory proposes that those with poor self-esteem seek out excessive reassurance from their social group, which can create conflict and lead to depression (Joiner, Alfano, & Metalsky, 1992). And lack of support in crisis coupled with low self-esteem greatly increases the risk of developing depression (Brown *et al*., 1986). Regardless of the exact mechanisms at hand, these variations in self-esteem influence an individual's vulnerability for depression as well as maintaining depressive symptomatology. Consequently, improving self-esteem may be an important treatment goal for ameliorating depression and related disorders.

One of the challenges in improving self-esteem, however, is the difficulty in accurately measuring it. Self-esteem, like symptoms of mood or suicidal thoughts, is primarily assessed exclusively via patient self-report. Concerns of the meaningfulness of answers on self-reports are widely recognized for healthy individuals (Kagan, 2007), and they are magnified in patients with

MDD and other internalizing disorders. Negative self-perception and negative schema are core constructs of MDD (Beck, 1987), and both create an overemphasis on the negative aspects of one's life (Wenze 2012). Additionally, a pressure may exist for patients to report inaccurate information on some symptoms for several motives including concern for the reactions from mental healthcare providers and social desirability. Social desirability does play a role in suicidal ideation (Linehan & Nielsen, 1981). Discrepancies exist between disclosures from self-reports and from clinical interviews in recent suicidal ideation (Kaplan *et al.*, 1994), and most measures of suicidality fail to be meaningfully predictive of future suicide attempts (Bryan & Rudd, 2006). These findings indicate that these important constructs, like self-esteem and suicidality, can be difficult to measure accurately.

Clinician rating scales and other collaborative methods of assessment (*e.g.* the calendar method) can only partially ameliorate the aforementioned limitations of self-reported measurement. Thorough and collaborative measurements such as the Timeline Follow-back method (Sobell, *et al*., 1979) and Life History Calendar method (Freedman, *et al.* 1988) do exist, and have been shown to be psychometrically sound for several psychological disorders including Substance Use Disorder (Sobell & Sobell, 1992), trauma history (Yoshihama *et al*., 2002), and even for general social and economic history (Belli, Shay, & Stafford, 2001). But these methods also rely on the patient's memory –albeit with prompts for a methodical reflection by the assessor– and can often be time consuming, which restricts their use. These methods do bypass some biases and heuristics by using salient events as anchors; however, at the same time, they are influenced by others issues like the anchoring and availability heuristics. For example, individuals remembering that they felt very worthless on Christmas Day would make them more likely to error in favor of recalling feeling worthless during the holiday season. Moreover, these methods also rely on the willingness of the patient to disclose this information and their ability to do so meaningfully, just as self-report methods do. Methods that are solely based on observations from the clinician may be affected by similar errors in the clinician's judgment like the confirmation bias (seeking out confirmatory information and disregarding alternatives; Nickerson, 1998), especially if assessors are not blinded to the patient's state and well-trained in the assessment.

Self-esteem is one such concept that may be difficult to measure meaningfully for the aforementioned reasons. Memory is mutable and imprecise (Schacter, 2002; Bradburn *et al.*, 1987). Several biases and heuristics exist that can alter self-reports. Tversky and Kahneman (1974) identified two such heuristics: anchoring (over-reliance on the first piece of datum as a starting point) and availability (weighing information heavily simply because it can be recalled more readily). Biases like the serial position effect (recalling data at the beginning and end of a timeframe more accurately than data in the middle) and peak-end rule (overemphasizing the most intense moment and the final moment) will alter self-reports as well (Shiffman, Stone, & Hufford, 2008; Hurlstone, Hitch, & Baddeley, 2014). These errors and shortcuts cast doubt onto one's ability to meaningfully and accurately aggregate information over time. These shortcomings also extend into self-reports of self-esteem, and are furthered by the negative self-perception that characterizes MDD. The negative self-perception predisposes one to remember negative stimuli more readily (Gotlib *et al.*, 2004). For example, since a patient could quickly remember several days in which he felt worthless and none in which he did not, he assumed that he had felt this way for almost the entire month.  Not only are patients' memory's affected by the typical heuristics and biases associated with retrospective self-report, but patients' memories of their emotional states are also affected by the biases associated with MDD and the interaction between the two.

People report differently on emotions when they are actually experiencing them versus when they are not experiencing them (e.g., Stone et al., 1998). Several theories (Conner & Barrett, 2012; Robinson & Clore, 2002) have emerged to help to explain this discrepancy seen between ecological momentary assessment (EMA) and retrospective self-report. One line of research has begun to divide recall and memory into three constructs: experiences, memories, and beliefs (Singer & Salovey, 1993; Wilson, 2009, Kahneman & Riis, 2005). The remembering self is assessed through retrospective self-report measures; the experiencing self is closely related to day-to-day experiences and measured by EMA; the believing self, measured by trait assessment, serves as scaffolding for one's interpretation of incoming information and "for anticipating and forecasting the future" (Conner & Barrett, 2012). This tripartition can cast doubt on the meaningfulness of retrospective report because the experiencing self is "functionally and neuroanatomically different from the 'remembering' and 'believing' selves measured through retrospective and trait questionnaires" (Conner & Barrett, 2012, p. 327). A different neural circuit activates when one encodes in-the-moment information about an event when compared to when one later recalls how one was feeling in that moment. The responses on retrospective self-reports may therefore align more closely with what one believes about how one felt at a particular moment than how one actually felt at that point in time, which calls into question the meaningfulness of retrospectively recalling how one felt at a point in the past. This discrepancy in self-reporting retrospective information becomes more of a problem in clinical populations; for example, Beck (1987) suggests that a depressed patient's negative self-schema greatly influences his perceptions of his own life because the believing self affects the interpretation of information.  This biasing indicates that retrospective self-report of emotions may not be completely accurate and may even be different from how one was actually feeling in the moment.

Robinson and Clore (2002) provide a similar explanation for the disconnection between retrospective self-report and EMA. They suggest that when asked to report on feelings that are not currently experienced, the reporter will access his beliefs about his affect at the time rather than accessing the past affective state itself. In other words, they hypothesize that, when completing retrospective self-report, one's remembering self will come online when thinking about the situation instead of one's experiencing self. For these reasons, the interplay of these three memory structures may lead to inaccuracies during retrospective self-report of feelings and emotions.

One way to address the disconnect between how individuals feel at the time they experience something and how they retrospectively report on those experiences is the use of implicit assessments. While not completely free of bias, implicit assessments can provide insights into constructs of interest that are not subject to the biases outlined above.  The negativity bias (paying more attention to negative stimuli as well as negative events affecting psychological states more than positive events) seen in MDD (McCabe & Gotlib, 1993; Beevers & Carver, 2003), threat attention in PTSD (Fani, *et al*., 2012; Foa, *et al*., 1991) and weight bias in Anorexia Nervosa (Dobson and Donzois, 2004) have all been successfully examined using implicit measurements. A successful implicit task – the Implicit Association Task (IAT; Greenwald, McGhee & Schwartz, 1998) uses button press reaction times to measure one's implicit associations. Performance on the IAT has been shown to be stable over time (Greenwald & Farnham, 2000), and, furthermore, the IAT may buffer against social desirability. When participants with anxiety disorders were asked to present a positive impression, they succeeded at masking their levels of anxiety in a retrospective self-report, but were unable to do so on an IAT (Egloff & Schmukle, 2002).

The IAT has been used successfully to study questions with high social desirability like racial prejudice and suicidal thoughts, and has been shown to aid in prediction of these traits (Greenwald, *et al.*, 2009). Participants will not typically self-disclose racial biases on self-report measurements; however, time and again, studies using the IAT demonstrate a clear pattern of rating one's own racial or ethnic group as the most positive and the other groups in a particular order (Whites> Asians> Blacks> Hispanics) (Axt, Ebersole & Nosek, 2014; Howell, Gaither & Ratliff, 2014). This finding of a set pattern extends to religious affiliation and age groups as well. After participants completed an IAT measuring implicit racial associations, Howell and colleagues (2014) questioned participants about these implicit racial biases and most participants became defensive and blamed the task for producing incorrect results, confirming the strong social demand to present in a particular way (*i.e.* not racist). However, because the IAT has strong retest reliability, and evidence of some immunity to "faking good" exists, it seems as though the IAT can accurately assess some implicit core beliefs that may not be readily expressed due to social desirability and other external pressures to present in a particularly way.

With regards to suicidal behavior, more-recent work by Nock and colleagues (2010) has demonstrated the incremental validity of the IAT using death-related stimuli to predict attempts of suicide. In this study, the IAT provided strong incremental validity over all other factors including one's history of suicidal behavior. Furthermore, in this study, the IAT demonstrated strong predictive validity – incrementally more predictive than both clinician and patient ratings of the risk of suicidal behavior (Nock, et al., 2010). This finding demonstrates that such an implicit measurement of a construct that is difficult to assess can have practical clinical utility in the treatment of psychological disorders like MDD.

However, despite promising results in the area of suicidality, the implicit tasks have been less successful in measuring constructs related to symptoms of worthlessness and self-esteem. Few studies demonstrating correlations between explicit and implicit measurements of self-esteem have been reported (Glashouwer & de Jong, 2010; Risch *et al.*, 2010; Franck, De Raedt, & De Houwer, 2007; Buhrmester *et al.*, 2011). At the very least, modest correlation with explicit measurements of self-report would be expected. Mild correlations exist between EMA and explicit measurements, which are expected since the three types memory outlined by Conner and Barrett (2012) are distinct but related. However, results from studies assessing self-esteem suggest that implicit and explicit measurements of self-esteem assess distinct concepts that only correlate mildly with each other (Greenwald & Farnham, 2000; Bosson, Swann Jr., & Pennebaker, 2000). Furthermore, several studies using depressed and formerly depressed patients found no meaningful correlation between task performance and mood after inducing a negative mood (Franck, De Raedt, & De Houwer, 2008; Gemar, Segal, Sagrati, & Kennedy, 2001). Other studies have failed to correlate with explicit measures of self-esteem or even to depression symptoms specifically (Lemmens *et al*., 2014; Bosson, Swann, & Pennebaker, 2000). Lastly, a meta-analysis indicated that the IAT correlates weakly with explicit reports of self-esteem and fails to correlate meaningfully with covariates typically associated with changes in self-esteem like depression and well-being (Buhrmester, *et al*., 2011). These mixed results indicate that implicit assessments of self-esteem may not be assessing the construct accurately or wholly.

Several explanations for these differences have been put forth. One suggestion is that self-esteem is not a unitary construct (Cunningham, Preacher, & Banaji, 2001). Indeed, many self-report questionnaires have subscales for different domains of self-esteem. It is not unreasonable to imagine individuals who think of themselves as hardworking but who hate the way they looks, or, conversely, individuals who think of themselves as attractive but offering little else. Moreover, many of the studies looking at implicit measurements of self-esteem "put the cart before the horse" by attempting to validate the task using depressed and/or remitted populations. While depression and self-esteem are correlated (Buhrmester *et al*., 2011), they are not isomorphic constructs. Depression may affect sub-constructs of self-esteem differentially and not necessarily in the same way for each individual. This idea coupled with the heterogeneity in the construct of self-esteem may help to account for the lack of consistency in the studies that attempt to examine self-esteem implicitly. This lack of consistency indicates that alternative methods of assessment are needed.

An advantage to using a different approach – fear conditioning – to assess self-esteem implicitly is the ability to leverage the well-characterized circuitry of classical fear learning and extinction. The neurobiology of fear acquisition and extinction has been heavily researched using a translational neuroscience approach (LeDoux, 1996; Maren & Quirk, 2004; Delgado, Olsson, & Phelps, 2006), which has demonstrated the importance of the prefrontal cortex in impacting the fear learning that occurs in the amygdala. Work from animal models indicated that sensory information pertaining to stimuli is projected from the thalamus to the lateral nucleus of the amygdala (Figure 1). These signals are carried, via the central nucleus of the amygdala, to the regions in the hypothalamus and brainstem that mediate the conditioned fear response (Maren & Quirk, 2004; Delgado, Olsson, & Phelps, 2006). Importantly, this neural circuitry seems not only to exist but also to be critical for fear conditioning across species (LeDoux, 1996). In animal models, the prelimbic prefrontal cortex (PFC) has been heavily implicated in the consolidation of memories of fear learning. Brain derived neurotrophic factor (BDNF) – a protein abundant in the prelimbic PFC – and many other brain regions – is involved in synaptic plasticity (a key neurobiological mechanism of memory consolidation, among other functions). Mice that have been altered to have no BDNF in the prelimbic PFC demonstrate impaired fear conditioning (Choi et al., 2010). This impairment in fear conditioning can be rescued by means of the administration of a BDNF analog (Choi et al., 2010), which indicates that synaptic plasticity in the prelimbic PFC is necessary for fear conditioning. Furthermore, stimulation of the medial PFC (mPFC) has been shown to be important in extinction learning, as lesions of the mPFC inhibit extinction learning (Phelps *et al*., 2004). Conversely, stimulation of the mPFC (specifically the infralimbic portion) aids extinction memory (for overview, see Quirk, Garcia, & González-Lima, 2006). More recently, investigators have found expected results when investigating fear conditioning and subsequent extinction of that association by means of neuroimaging techniques in humans (for overview, see Delgado *et al*., 2008). These findings are consistent across studies and are based upon a well-established cross-species literature, which demonstrate both that the amygdala is the central location for fear learning and that the mPFC can modulate that learning.

In contrast, the neurological correlates of the IAT are not wholly understood, as few studies to date have examined them. Discovering this neural circuitry is further complicated since training animals in the IAT may prove difficult, if not impossible. Several human subjects studies have investigated differential activation during "compatible" versus "incompatible" trials, and several prefrontal cortical regions emerge from those contrasts. These areas span the anterior cingulate cortex (ACC) as well as much of the frontal lobe: the ventro- and dorso-lateral prefrontal cortex (vlPFC and dlPFC) and the orbitofrontal cortex (Beer *et al*., 2008; Luo *et al*., 2006; Chee *et al*., 2000). All of these regions are associated with inhibitory control (Aron *et al*.,

2004) and are also seen in other effortful processing tasks like the N-back task (Owen, *et al.*, 2005).

Furthermore, the areas of differential activation may depend largely upon the stimuli type used in the task. For example, one study using a morality-based IAT found differential activation in the amygdala (typically active in both moral reasoning and automatic processing) and vlPFC (Luo *et al.*, 2006). Another using natural objects (insects vs. flowers with pleasant vs. unpleasant) as the stimuli only found changes in activation in the dlPFC (Chee *et al*., 2000). Beer and colleagues (2008) used the IAT to assess implicit racial attitudes and discovered differences in activation in the striatum (caudate), insular cortex, and orbitofrontal cortex. Studies using clinical populations like Substance Use Disorder using both alcohol- and marijuana-based IATs saw differences in frontal and striatal regions also (Ames *et al*. 2013; Ames *et al*., 2014). However, these studies simply demonstrate that areas of cognitive control and automatic processing are involved when one has to sort the stimuli during an IAT. The activation of only areas associated with cognitive control is problematic to the underlying hypothesis of the IAT: that individual's core beliefs drive the differential response rate between compatible and incompatible trials. Yet, there is no indication from any of the neuroimaging studies in humans that indicates any differences in metacognitive comparison while completing the IAT. Simply stated, the non-specificity of these findings makes transitioning the IAT into neuroimaging paradigms more difficult.

The experimental paradigm described below utilized a different approach – classical fear conditioning – in an attempt to bypass several of the perceived limitations of the IAT in measuring self-esteem. Classical conditioning is an effective paradigm (Beckers, *et al*., 2013; Bouton, 2007), and it is understood that the aversive response will transfer to the conditioned stimulus after acquisition of the association (Hermans, *et al.*, 2002). In order to make the self-related stimulus as representative of oneself as possible, we used a photograph of each participant as a conditioned stimulus. Additionally, an acoustic startle, which is also well researched and widely used in fear conditioning (*e.g.* Davis, *et al.*, 1982), served as the unconditioned stimulus. The conditioned response is pupil dilation; pupillometry is an effective and well-validated metric for physiological arousal (Bradley, *et al.*, 2008), and is effective in fear-conditioning paradigms (Reinhard, Lachnist, & Konig, 2006).

## Overview of Current Studies

A paucity of validated implicit measurements of self-esteem exists. The present research, which sought to fill this gap, has two aims. The first aim is to determine the feasibility of using a fear-conditioned pupil response to self-related stimuli as an implicit measurement. The second aim is to investigate if any meaningful correlations of the conditioned pupil response to self-reported measurements of self-esteem exists. The following data are results collected from three studies demonstrating the development and validation of this aversive conditioning paradigm. Study 1 was the first attempt in development of the paradigm. Study 2 modified the original task design to enhance detection of the conditioned response (pupil dilation). With the modified design validated, Study 3 investigated our second aim: the relationship between the conditioned pupil response and self-reports of self-esteem and related constructs.

## General Methods

This study utilized an aversive conditioning paradigm with faces as conditioned stimuli in order to investigate the role one’s self-esteem has on modulating autonomic responses (pupil dilation) to self-related imagery. The unconditioned stimulus (UCS) used in this task was an unpleasant noise delivered binaurally through headphones. The innate acoustic startle reflex

(Koch, 1999) is the unconditioned response (UCR). The conditioned stimuli (CSs) were photographs of faces. The UCS onset was the final second of CS presentation and the aversive noise co-terminated with the CS. The conditioned response (CR) was pupil dilation in response to the faces that were paired with the UCS. Pupil area has been shown to be a reliable, unobtrusive measure of autonomic response (Bradley, *et al.*, 2008; Partala & Surakka, 2003; Goldwater, 1972).

Upon arrival to the laboratory, participants were given a brief overview of the study and provided informed consent. After consent, each participant completed self-report questionnaires (see below). The experimenter took a photograph (CS-Self) of all participants in Study 1 and Study 3, but not those in Study 2 since the CS-Self was not used in that paradigm. This CS-Self was paired with the UCS in some studies in order to serve as a CS+. After all self-report questionnaires were completed, participants were taken to a secondary location for the computerized conditioning task. Participants were first habituated to the all CSs by viewing them each twice for 6 seconds in a fixed random order on a computer monitor. Acquisition and extinction blocks followed. Afterwards, participants underwent a positively-valenced counter-conditioning task involving rewards that were paired with self-related imagery in order to offset any deleterious effects of the CS-Self being paired with the UCS. Each block of the conditioning task occurred without the experimenter present in the room. The room lights were also left on during all tasks. After the conditioning tasks, participants were brought back to the initial location where they completed a debriefing questionnaire. After doing so, study staff debriefed each participant and then dismissed him or her. The Emory University Institutional Review Board approved all study procedures. The study participants were undergraduate students enrolled in Introductory Psychology at Emory University who, as a part of the curriculum, are required to participate in research studies.

**Self-Report Questionnaires**

Several self-report measures were used to assess concepts of self-esteem: the Neff Self-Compassion Scale (NSCS) (Neff, 2003); the Rosenberg Self-Esteem Scale (RSES) (Robins et al., 2001); and the State Self-Esteem Scale (SSES) (Heatherton & Polivy, 1991). The State-Trait Anxiety Inventory (STAI) (Spielberger et al., 1970); and the Perceived Stress Scale (PSS) (Cohen, Kamarck, & Mermelstein 1983) were used to measure stress and anxiety. These two constructs represented possible confounds since pupil dilation generally correlates with autonomic arousal, which can also be induced by stress. The Center for Epidemiological Studies – Depression (CES-D) (Radloff, 1977) was included to measure depressive symptoms, as negative self-concept is a core symptom of depression. The Positive and Negative Affect Schedule (PANAS-Now) (Watson et al., 1988) was included also. Lastly, participants completed a simple demographics questionnaire developed by our laboratory for research purposes.

**Stimuli**

The UCS was an unpleasant, aversive noise presented binaurally through Apple Earpods at between 80-85dB (set prior to participant arrival by means of a decibel meter). The non-self CS+ face and CS– faces were matched as closely as possible to each participant's self-identified gender and ethnicity, and were constant in each of the eight possible gender-race combinations. The photographs of faces were taken from publicly available facial stimuli databases (Ebner et al., 2008; Minear & Park, 2004; Thomaz & Giraldi, 2010). The photograph of the participant was taken by the experimenter using the Photobooth application on a MacBook Air. All CSs were edited to be grayscale and to have the same luminance properties using Seashore and GIMP (photo-editing programs). The UCS was paired with both CS+ faces (self and non-self) for each trial during acquisition, but never paired with the CS–.

## Eye-tracking

An Eyelink 1000 Plus –a dark-pupil-based eye-tracker– was used for all study participants. The participants were familiarized with the eye-tracking procedure prior to tasks, and the eye-tracker was calibrated to their eyes prior to each block in the conditioning task. The participant's right eye was tracked.

## Habituation

Habituation to all CSs presented during the acquisition and extinction blocks was performed first. The participants passively viewed a presentation of each CS individually for 6000ms in a fixed random order such that each CS was displayed on-screen twice.

## Acquisition

Following habituation, acquisition of the UCS-CS+ pairing occurred. Each block of acquisition trials consisted of eight trials of each CS in a fixed random order, totaling 24 blocks in all for Study 1 and Study 3, and 16 blocks for Study 2. After each block, the participants completed a contingency awareness rating (likelihood of hearing noise) for each CS by means of a computerized visual analog scale.

## Extinction

Extinction followed the acquisition phase. Each block of extinction consisted of eight trials of each CS in a fixed random order, in which the UCS was never presented. All participants underwent two blocks of extinction.

## Positively-valenced counter-conditioning (PVCC)

Following extinction, participants completed a positively-valenced counter-conditioning task to offset any potential effects of learning an association of self-related imagery with an aversive stimulus. During PVCC, each participant was shown three pairs of images – one pair contained the CS-Self, another pair used the CS– from the acquisition and extinction phases, and the third was comprised of two photographs of faces not previously viewed but from the same databases.

Each of the image pairs was associated with a given outcome: gaining $10 or $0; losing $10 or $0; or a neutral response (gray square or nothing). For each of the 96 trials (32 per condition), the pairs were presented in a fixed random order after a fixation cross, with one stimulus on the left and the other on the right side of the screen (counterbalanced). The participants were instructed to choose the left or right stimulus by pressing the “s” or “l” key respectively. The choice was then circled in red and the participants were given feedback on the outcome of their choice. For this task, the CS-Self was always in the “Gain” pair, associated with an 80% chance of a monetary gain. Although participants were instructed to try and earn as much as possible, they were told that all monetary reward for this task was hypothetical.

**Debriefing**

Participants completed a debriefing questionnaire. Afterward doing so, because the study employed minor deception, participants were told of the deception and the reasons for using it. They were then given the opportunity to ask any questions before study completion.

**Data processing**

After completing the study, participants’ pupillometry data were extracted through Eyelink Data Viewer – Eyelink’s propriety analytic software. In order to minimize variance, trial data was filtered to include information from 2000ms (after the initial light reflex) to just prior to the start of the UCS (approximately 3000ms total). Data was also only extracted from fixations inside an interest area set to correspond to the location of the CSs on-screen. Data for each trial was then converted into a percentage of the average pupil area (Ling & Han, 2009; Kaiser, 1989) for each block, on a block-by-block basis, in order to reduce the effects of a large linear habituation trend that occurred throughout the conditioning task. These percentages were then averaged over each block for each of the CSs. In order to ensure no major variations (such as a participant viewing one face while ignoring others), the interest area dwell time and number of fixations for each trial were also analyzed.

## Study Design

### Study 1

Study 1 employed our first task design: three CS (including CS-Self) presented individually on a black background (Figure 2). Forty-eight participants completed Study 1. The participants completed several self-reports prior to the conditioning tasks: demographics form, STAI, NSCS, RSES, CES-D, PANAS-Now, and PSS. This study employed two CS+s and one CS–, with the CS-Self serving as a CS+. After an inter-trial interval of either 4500, 5000 or 5500ms (fixed random order), a fixation cross appeared on a black background for 2000, 2500 or 3000ms (fixed random order), and participants were instructed to look at the fixation cross. Then, a CS was presented individually on-screen for 6000ms. After 5000ms, the UCS onset occurred, and both terminated 1000ms later. The presentation of the CS was followed by another inter-trial interval, and so on, for all CS presentations. Participants moved onto the extinction phase when either the participant rated each CS+ at ≥90% and the CS– at ≤10% or they had completed a total of four blocks of acquisition. These participants completed PVCC using the CS-Self to offset any deleterious effects of the UCS-to-CS-Self pairing.

### Study 2

Study 2 utilized a simpler design with two CSs, neither of which was the CS-Self. The two CSs used during this study were the race- and gender-matched CS+ and CS–. Instead of being black, the background was adjusted to be a grayscale value that matched average luminance of all of the CSs. Fixation crosses were no longer presented as well (Figure 3). The inter-trial interval (fixed random order of 3000, 3500 or 4000) was also shortened to accommodate the removal of the fixation cross. These changes were made after further consultation with peers in order to maximize the signal from the pupil response (modeled after Bradley *et al.*, 2008). Eleven participants completed this study. The participants only completed the demographics questionnaire prior to the conditioning tasks. Each participant completed habituation, at least two blocks of acquisition (regardless of the contingency awareness ratings after the first block of acquisition) and extinction.

### Study 3

In Study 3, the grey background was utilized again, and the fixation cross was also not presented. However, like Study 1, Study 3 presented participants with all three CSs: CS-Self, CS+ and CS–. The conditioning tasks in Study 3 (Figure 4) were identical to those outlined in Study 2. Thirty-seven participants completed Study 3. The participants completed the demographics form, STAI, NSCS, RSES, CES-D, PANAS-Now, PSS, and SSES prior to conditioning tasks. All subjects completed at least two blocks of acquisition trials, and all completed two blocks of extinction trials. After extinction, participants completed PVCC because the UCS was paired with the CS-Self.

## Results

### Study 1

Of the forty-eight participants who completed Study 1, fourteen participants were excluded for not meeting contingency awareness criteria by the end of the second block of acquisition. Of the thirty-eight remaining participants, ten were excluded due to an artifact in the eye-tracking data. Four others were also excluded: one participant reported having issues staying awake and diligent during the task, another's data was missing, and two others had issues when calibrating the eye tracker, leaving twenty participants. Of those participants, all completed at least one block of acquisition, but only nine completed the second block of acquisition, since eleven met the contingency awareness rating thresholds after the first block of acquisition. The mean age of this group was 18.95 (SD= 1.28). Seventeen participants identified as female (85%) and three as male. Demographically, these participants were predominantly White (10; 50%) and Asian (8; 40%) with one identifying as Black (5%), and one who did not identify a race. Nineteen of the twenty (95%) participants were right handed.

**Stimuli Ratings and Contingency Awareness.** The average rating of the unpleasantness of the noise (via debriefing questionnaire) was 6.33 (SD= 2.70), with 10 being completely unpleasant. The average rating of the likability of the photograph taken of each participant was 4.28 (SD= 2.20). The contingency awareness ratings after the first acquisition block were 86.16 (SD= 20.77) for CS+, 91.37 (SD= 21.67) for the CS-Self, and 17.42 (SD= 28.23) for CS– [n=19]. After the second acquisition block, the ratings were 99.14 (SD= 2.27) for CS+, 99.57 (SD= 1.13) for the CS-Self, and 1.71 (SD= 2.98) for the CS– [n=7].

**Pupillometry.** A 2x2x3 repeated measures ANOVA was performed using two epochs (acquisition and extinction), two blocks, and the three CS-types. Main effects of both epoch and block were significant ($F_{1,6}$ = 11.151, p=0.016; $F_{1,6}$ = 10.889, p =0.016), while the main effect of CS-type was not ($F_{2,5}$ = 0.799, p=0.500). The only significant interaction was that of epoch and block ($F_{1,6}$ = 8.192, p =0.029). These results indicate that while participants' explicit ratings of contingency awareness of the CS-UCS pairings, this conditioning was not demonstrated in pupillometry data (Figure 5). No significant differences emerged when analyzing just the data from the CS+ and CS– during acquisition in a repeated-measures ANOVA (main effect of CS-type: $F_{1,6}$ = .608, p=0.465; interaction of CS-type and block: $F_{1,6}$ =5.831, p= 0.052). This lack of expected results indicated that the study design needed amending because the anticipated conditioned responses were not borne out by the data.

**Study 2**

Study 2 was performed for two reasons. First, we wanted to determine the feasibility of using faces as CS in a classical conditioning paradigm because of the lack of differential pupil responses (CR) to an aversive noise demonstrated by Study 1 participants. Second, we wanted to ascertain whether or not the lack of a CR was due to the presentation of the CSs on a black background (which caused a large shift in luminance when the CSs were presented), and if using a luminance-matched gray background could ameliorate this issue. Eleven participants completed Study 2. The mean age was 19.00(1.05). Four identified as Asian (36.36%), four as White, and three (27.27%) as Black. 10 identified as female (90.91%), and one as male. Ten of the eleven participants were right handed (90.1%). These participants did not differ significantly from those in Study 1 in age ($t_{29}$ =0.091, p=0.928), gender ($\chi^2$,1= 0.220, p= 0.639), race ($\chi^2$,3 = 3.595, p=0.309), or handedness ($\chi^2$,1 = 1.787, p=0.181).

**Stimuli Ratings and Contingency Awareness.** The average rating of the unpleasantness of the noise (via debriefing questionnaire) was 7.55(2.88), with ten being completely unpleasant. Analysis indicated that Study 2 participants did not find the noise quite as unpleasant as those from Study 1 ($t_{27}$ = -1.145, p=0.262). After the first block of acquisition, the contingency awareness ratings were 89.36 (SD= 15.32) for CS+ and 14.55 (SD= 25.304) for CS– [n = 11]. By the end of the second block of acquisition, they were 97 (SD= 3.703) and 4.75 (SD= 5.203) respectively [n = 8]. No significant differences existed at either time point between the contingency awareness ratings for either CS+ ($t_{28}$ = -4.45, p=0.660; $t_{28}$ = 1.325, p=0.208) or CS– ($t_{13}$ = .262, p=0.796; $t_{13}$ = -1.357, p=0.198) when compared to Study 1 participants' contingency awareness ratings. This finding indicated that participants were learning the pairing at approximately the same rate in both of the studies.

**Pupillometry.** The pupillometry data provided evidence for the acquisition and subsequent extinction of the conditioned response using two, non-self faces. T-tests were performed due to the small sample size. Participants' pupils dilated significantly more to the CS+ than the CS– during the first ($t_{10}$=3.74, p=0.004) and second ($t_8$=3.38, p=0.010) blocks of acquisition, correcting for multiple comparisons. During the first block of extinction, this effect was reversed but not significant ($t_{10}$= -1.66, p=0.127). This trend corresponded with the participants' reporting that they expected the UCS to be paired with the CS– once they realized it was no longer paired with the CS+. No significant difference occurred in the second block of extinction either ($t_{10}$=1.39, p=0.195) (Figure 6). These participants explicitly reported awareness of the conditioning paradigm (since this awareness was used as a determinate to move onto the extinction epoch) as well as the expected implicit conditioned response. To extend the feasibility of using facial stimuli as CS in a conditioning paradigm, Study 3 added the CS–Self as a second CS+ in order to determine if the inclusion of self-related imagery affects the participants' abilities to acquire the conditioned response.

## Study 3

Thirty-seven participants completed Study 3. Five participants were excluded for not meeting contingency awareness criteria by the end of the second block of acquisition, and three others were excluded due to artifacts in the eye-tracking data. The mean age was 19.21 (SD= 1.207) and 20 of 29 were female (69%). Twelve identified as White (41.4%), ten as Asian (34.5%), four as Black (13.8%), two as American Indian (6.9%), and one as more than one race (3.4%). Additionally, twenty-seven participants (93.1%) were right-handed and two were left-handed. Because no significant differences existed in the demographic characteristics of Studies 1 and 2, and because they are samples drawn from the same population, we collapsed across Study 1 and Study 2 participants to create a larger reference group in order to increase power when analyzing the demographic characteristics of Study 3. Study 3 participants did not differ significantly from those in Studies 1 and 2 in age ($t_{58}$ = -0.880, p=0.382), gender ($\chi^2$,1= 2.902, p= 0.088), race ($\chi^2$,5 = 4.274, p= 0.511) or handedness ($\chi^2$,1 = 0.388, p=0.533).

**Stimuli Ratings and Contingency Awareness.** Study 3 participants also did not differ significantly from the larger reference group in the average subjective unpleasantness of the UCS ($t_{56}$= 0.333, p=0.740). These participants rated the UCS as 6.55(2.73) out of ten for unpleasantness. They also did not differ significantly ($t_{45}$ =0.349, p= 0.729) from participants from Studies 1 in self-reported liking of the photograph of themselves, which they rated, on average, as 4.03 (SD= 2.40) out of ten for likeability. The contingency awareness ratings after the first block of acquisition were as follows: CS+ 80.97 (SD= 20.21); CS-Self 80.31 (SD= 23.75); CS– 20.66 (SD= 28.97). After the second block of acquisition, the contingency awareness ratings were: CS+ 99.00 (SD= 3.22); CS-Self– 98.07 (SD= 5.45); CS– 1.38 (SD= 4.37). After both of these blocks, contingency awareness ratings were not significantly different from the combined reference group for CS+ ($t_{57}$ = 1.256, p=0.214; $t_{42}$ = -0.978, p=0.334); for

CS-Self ($t_{46}$ = 1.632, p=0.109; $t_{34}$ = 0.718, p=0.478); or for CS– ($t_{57}$ = -0.582, p=0.563; $t_{42}$ = 1.397, p=0.170).

**Pupillometry.** A 2x2x3 repeated measures ANOVA was performed using two epochs (acquisition and extinction), two blocks, and three CS-types. No main effects existed for either epoch or block ($F_{1,26}$= 0.154, p= 0.698 and $F_{1,26}$=1.047, p= 0.316). The assumption of sphericity was violated for CS-type, but the Greenhouse-Geisser corrected main effect of CS-type was still significant ($F_{1.460,37.97}$ = 7.565, p=0.004). Block*CS-type was the only significant interaction ($F_{2,25}$= 5.926, p =0.008). Visual inspection of the interaction indicated that CS– decreased from Block 1 to Block 2, while CS+ and CS-Self stayed the same or increased (Figure 7). To investigate this further, paired-samples t-tests were run to compare average pupil area for the three CSs during the second block of acquisition and the second block of extinction. Participants' pupils dilated more to the CS-Self than CS– ($t_{28}$ = 2.711, p=0.011), and more to the CS+ than CS– ($t_{28}$ = 2.857, p=0.008). However, no significant difference existed between the CS+ and CS-Self conditions ($t_{28}$ = 1.198, p=0.214), which indicated that participants did not differentiate between the two CSs paired with the UCS. The same pattern held true for the second block of extinction (CS+ v CS–: $t_{27}$ = 2.799, p=0.009; CS-Self v CS–: $t_{27}$ = 4.123, p=0.00032; CS-Self v CS+: $t_{27}$ = 1.612, p=0.119). The results of these t-tests survive a Holm-Bonferroni correction for multiple comparisons (Holm, 1979). These findings indicate that participants explicitly (via contingency awareness ratings) and implicitly (via conditioned pupil response) learned the conditioning pairings.

To increase the ability to detect meaningful trends throughout the data, a mixed-effects multi-level general linear modeling approach was employed using STATA 14. This approach was chosen for several reasons. First, this modeling would be able to better control for the large linear decreases in pupil area exhibited by participants over the course of each of the blocks and throughout the conditioning overall. Secondly, this approach increased the statistical power of the analyses by using data from each trial instead of collapsing over trials (*i.e.* averaging for each CS-type) in each block. The model included epoch, block, trial CS-type, age, gender, and debriefing ratings of both the UCS unpleasantness and the likeability of the participant's photograph as independent variables; the transformed pupil data was the dependent variable. Age and gender did not have any significant main effects, and were not significantly interacting with any other independent variables, so they were removed to simplify the model.

The model was run with several self-report measures included separately as independent variables, and the only one that accounted for any significant proportion of the variance in conditioned pupil response was the Neff Self-Compassion Scale (NSCS). In this model (epoch, block, CS-type, UCS rating, CS-Self rating, & NSCS score), with CS-Self as the reference group, a significant main effect of CS-type was found for both CS– (z= -2.37, p= 0.018) and CS+ (z= -3.10, p= 0.002). This finding indicated that the CS-Self was causing larger pupil dilations than the CS– and CS+ throughout the four blocks. Additionally, interaction of NSCS score and CS-type was also significant for CS-Self compared against CS– (z= -2.12, p= 0.034). This interaction may indicate that the underlying constructs assessed by the self-compassion questionnaire interfere with learning the conditioned response. To investigate the effects of self-compassion (via NSCS score) further, we ran the model for each epoch individually. When doing so, the interaction between NSCS score and CS-type (for CS-Self vs. CS–) accounted for a significant proportion of the variance during acquisition (z= -2.23, p= 0.026), but not extinction (z= 0.83, p= 0.409). To visualize the interaction during acquisition, each participant's NSCS score was transformed into a categorical variable and binned into quintiles in order to get two extremes and a midpoint. Then, the estimated margins from the interaction of quintiles with CS-type were graphed for the highest, middle and lowest quintile (Figure 8). The plot of these three bins depicts different trends in conditioned response to the CS-Self between the highest and lowest quintiles, indicating that self-compassion may be modulating the conditioning.

To further understand how the NSCS is modulating the learning, we ran the model with each of the six NSCS subscales individually during acquisition and discovered that the NSCS*CS-type interaction for CS-Self vs. CS– was driven by just two of the six subscales: Self-Kindness (z= -2.54, p= 0.011) and Self-Judgment (z= -2.01, p= 0.044). These two broad constructs map onto symptoms of depression –feelings of worthlessness and guilt– as well as the negative self-schema that may drive depressive symptoms. The interaction between CS-type and the other four subscales – Common Humanity, Over-Identified, Mindfulness and Isolation – did not account for any significant proportion of variance. The questions comprising the Self-Judgment and Self-Kindness subscales directly assess the one's criticalness of oneself, while the other subscales' items do not. The significant interaction with these two subscales indicates that being very self-critical may facilitate aversive conditioning to self-related imagery.

In order to investigate how self-report scores may relate to the differentiation in conditioned pupil responses to the CS+ and CS-Self (the two CSs paired with the UCS), we investigated those CS-types in more detail. When looking at the effect of NSCS scores on pupillometry results in the first block of acquisition in just CS+ vs. CS-Self, the main effect of NSCS was significant (z=2.06, p= 0.039) and the interaction of CS-type and NSCS score is also significant (z=-3.28, p= 0.001). This interaction remained significant when looking at the second block of acquisition as well (z=-2.21, p= 0.027). These findings did not reach significance in either block of extinction. Importantly, these analyses were not corrected for multiple comparisons. Only one test survived a Holm-Bonferroni correction for multiple comparisons: the interaction of NSCS score and CS-type when restricting the model to the first block of acquisition and comparing only CS+ with CS-Self.

## Discussion

Low self-esteem is a core symptom associated with risk, onset and maintenance of multiple psychological disorders. It is also a difficult symptom to measure accurately. These three studies were designed to address this challenge by determining the feasibility of using an aversive conditioning paradigm with self-related imagery in order to measure one's beliefs about oneself implicitly. Overall, the results from the studies herein indicate the feasibility of using self-related imagery as conditioned stimuli in such an experimental paradigm. Additionally, we have preliminarily evidence that this paradigm may be effective at measuring a specific sub-construct of self-compassion: the level of one's self-criticalness.

Self-esteem is a related to, but distinct from, self-compassion. Self-esteem, broadly, is how one feels about oneself. And this evaluation can change. For example, doing well on a test might increase one's self-esteem, but failing a test might decrease one's self-esteem; self-esteem is contingent on the responses received from one's environment. Self-compassion, on the other hand, is the extension of compassion toward oneself regardless of one's circumstances. This concept has been called several things in the past, including "unconditional positive regard" (Rogers & Dorfmann, 1951) and unconditional self-acceptance" (Ellis, 1977), but has recently been reintroduced as an important factor in psychological health (Neff, 2003).

Findings from the mixed effects modeling indicated that the self-compassion (as measured by the Neff Self-Compassion scale) might modulate the autonomic pupil response to self-related imagery in an aversive conditioning paradigm. Two subscales relating to self-judgment and self-kindness drove this effect. This finding is important because these two subscales parallel critical symptoms in depression associated with the negative self-schema that increases depressive symptoms. Investigators are currently exploring the role that self-esteem plays in the development of depression and focusing on self-esteem as a possible treatment target to help prevent and to treat depression. Our preliminary findings may be indicative that being highly self-critical may facilitate one's ability to learn a conditioning response of an aversive stimulus to self-related stimuli. This interpretation, however, must be made with caution for two important reasons. Firstly, no other self-report measures of self-esteem meaningfully correlated with the pupillometry data. Secondly, the analyses were not corrected for multiple comparisons. Several other self-report questionnaires and their subscales (like the Appearance subscale of the SSES, which is comprised of items involving satisfaction with one's appearance) measured similar constructs. One would expect that these could similarly modulate conditioning to self-related imagery, but no evidence of this relationship existed in our data. Although appropriate divergent validity is demonstrated in the absence of meaningful relationships with self-reports of anxiety and depression, some predicted correlations with relevant self-esteem measures are not apparent in this data. However, this failure is not unique to these findings, as few studies have successfully measured self-esteem using implicit measurements.

Additionally, the results from these studies indicated several methodological points to consider when developing such a task. First, our findings indicate that using a black background may obscure any differences in conditioned pupil response. The contingency awareness data from Study 1 indicate that participants were explicitly aware of the conditioning. However, the pupillometry results fail to show any meaningful distinctions between a face always paired with the UCS during acquisition and the face never paired with the UCS (Figure 5). When we used a luminance-matched gray background instead of a black one, participants demonstrated significant differences in conditioned pupil responses (Figures 6 & 7) during acquisition as expected. We hypothesize that the light reflex when switching from a solid black background to a grayscale image negated any meaningful differences in differential pupil dilation. Furthermore, a large habituation trend through all of the conditioning tasks complicated the interference. This hypothesis is supported by evidence of a clear conditioned response in Studies 2 and 3, even when controlling for both the presence of three conditioned stimuli (Figure 6) and a self-related CS+ (Figures 7). However, other studies have demonstrated clear conditioning responses to facial stimuli when presenting images on a black background, indicating that, while we were unable to get a conditioned response, it is certainly possible to do so.

Second, our findings emphasize the necessity of accounting for details that affect pupil dilation. Changes in luminance need to be minimized throughout the entire study. Since the pupil will naturally contract after fixating on a point, the window of time analyzed when assessing pupil response needs to be long enough to capture any meaningful differences. Finally, data from all three of these studies revealed a large effect of habituation within each block and through the experiment overall; this may mean that using the same UCS for all trials, as well as running large amounts of trials, may also interfere with the detection of significant differences in pupil response.

## Limitations

Several limitations of this study exist. The sample size is small and likely underpowered for higher-level statistical analysis like repeated measures ANOVAs to draw meaningful conclusions. Additionally, no corrections for multiple comparisons were made. Given the number of analyses run, it is unlikely that these effects would survive correction for multiple comparisons. Regardless, these findings do demonstrate the feasibility of using pupillometry responses in a conditioning paradigm as an implicit measurement of self-esteem. This paradigm may prove useful because of its correspondence with the cognitive-behavioral theory of depression in which negative self-schema cause incoming information to be interpreted negatively. This negative encoding bias moderately parallels conditioning: learning to associate self-related stimuli with one's negative self-concept further strengthens the automatic negative self-schema. Additionally, the use of a stimulus very close in approximation to oneself (*i.e.*, a still photograph of one's face) and the removal of some possible biases coming from the participant are both important steps in the development of an implicit measurement of self-esteem.

**Future Directions**

This line of research can be strengthened in several ways. First, the results need replicated. Just as with any preliminary finding, replication is needed before having confidence in these findings. Running another cohort of subjects through this paradigm will allow us to increase our statistical power for detecting an effect and enable us to confirm the findings that Self-Kindness and Self-Judgment interfere with conditioning self-related imagery with an aversive stimulus. Second, further exploration of the role that the self-related imagery has in the conditioning could shed light on the role of self-kindness in modulating that learning. The CS-Self in these experiments was always presented as a CS+; however, the self-related imagery could be broadly serving as a safety signal. Investigation of conditioning with this paradigm using the CS-Self as a CS– would provide insight into the broader role that self-related imagery has on conditioning. Third, because our goal is to develop an implicit measurement that is grounded in translational science, activation of the fear learning circuit (Figure 1) should be present while participants complete the conditioning tasks while simultaneous undergoing fMRI. Differential activation patterns to the three conditioned stimuli could provide more insight into the roles that self-kindness and the self-related imagery have in the learning process.

## References

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. Text Revision – Fourth Edition. Washington, D.C.: American Psychiatric Association; 2000

Ames, S.L., Grenard, J.L., He, Q., Stacy, A.W., Wong, S.W., Xiao, L., … & Bechara, A. (2014). Functional imaging of an alcohol-implicit association test (IAT). *Addiction biology, 19*(3), 467-481.

Ames, S. L., Grenard, J. L., Stacy, A. W., Xiao, L., He, Q., Wong, S. W., ... & Bechara, A. (2013). Functional imaging of implicit marijuana associations during performance on an implicit association test (IAT). *Behavioural brain research*, 256, 494-502.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in cognitive sciences, 8*(4), 170-177.

Axt, J. R., Ebersole, C. R., & Nosek, B. A. (2014). The rules of implicit evaluation by race, religion, and age. *Psychological science, 25*(9), 1804-15.

Beck, A. T. (1987). Cognitive models of depression. *The Journal of Cognitive Psychotherapy: An International Quarterly*, 1, 5-37.

Beckers, T., Krypotos, A. M., Boddez, Y., Effting, M., & Kindt, M. (2013). What's wrong with fear conditioning?. *Biological psychology, 92*(1), 90-96.

Beer, J. S., Stallen, M., Lombardo, M. V., Gonsalkorale, K., Cunningham, W. A., & Sherman, J. W. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *NeuroImage, 43*(4), 775-783.

Beevers C.G., & Carver C. (2003). Attentional Bias and Mood Persistence as Prospective Predictors of Dysphoria. *Cognitive Therapy and Research, 27*(6):619-37.

Belli, R. F., Shay, W. L., & Stafford, F. P. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public opinion quarterly, 65*(1), 45-74.

Bosson, J. K., Swann Jr., W.B., & Pennebaker, J.W. (2000). Stalking the perfect measure of implicit self-esteem: the blind men and the elephant revisited? *Journal of personality and social psychology, 79*(4), 631–643.

Bouton, M. E. (2007). Learning and behavior: A contemporary synthesis. Sinauer Associates.

Bradburn N., Rips L., Shevell S., (1987). Answering autobiographical questions: the impact of memory and inference on surveys. *Science*, 236:157–61

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602-607.

Brown, G. W., Andrews, B., Harris, T., Adler, Z., & Bridge, L. (1986). Social support, self-esteem and depression. *Psychological medicine, 16*(04), 813-831.

Bryan, C. J., & Rudd, M. D. (2006). Advances in the assessment of suicide risk. *Journal of clinical psychology, 62*(2), 185-200.

Buhrmester, M. D., Blanton, H., & Swann Jr., W. B. (2011). Implicit self-esteem: nature, measurement, and a new way forward. *Journal of personality and social psychology*, *100*(2), 365.

Centers for Disease Control, (2010). Current Depression Among Adults – United States, 2006 and 2008. *Morbidity and Mortality Weekly Report, 59*(38); 1229-1235.

Chee, M. W., Sriram, N., Soon, C. S., & Lee, K. M. (2000). Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. *Neuroreport, 11*(1), 135-140.

Choi, D. C., Maguschak, K. A., Ye, K., Jang, S. W., Myers, K. M., & Ressler, K. J. (2010). Prelimbic cortical BDNF is required for memory of learned fear but not extinction or innate fear. *Proceedings of the national academy of sciences, 107*(6), 2675-2680.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of health and social behavior*, 385-396.

Conner, T. S., & Barrett, L. F. (2012). Trends in ambulatory self-report: the role of momentary experience in psychosomatic medicine. *Psychosomatic medicine, 74*(4), 327-337.

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *Journal of consulting and clinical psychology, 76*(6), 909.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological science, 12*(2), 163-170.

Davis, M., Gendelman, D. S., Tischler, M. D., & Gendelman, P. M. (1982). A primary acoustic startle circuit: lesion and stimulation studies. *The Journal of neuroscience, 2*(6), 791-805.

Delgado, M. R., Nearing, K. I., LeDoux, J. E., & Phelps, E. A. (2008). Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. *Neuron, 59*(5), 829-838.

Delgado, M. R., Olsson, A., & Phelps, E. A. (2006). Extending animal models of fear conditioning to humans. *Biological psychology, 73*(1), 39-48.

Dobson, K. S., & Dozois, D. J. (2004). Attentional biases in eating disorders: A meta-analytic review of Stroop performance. *Clinical psychology review, 23*(8), 1001-1022.

Ebner NC, Riediger M, Lindenberger U. FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Germany: Max Planck Institute for Human Development; 2008. Unpublished manuscript.

Egloff, B.; Schmukle, S.C. (2002), Predictive Validity of an Implicit Association Test for Assessing Anxiety. *Journal of personality and social psychology*, 83: 1441–1455.

Ellis, A. (1977). Psychotherapy and the value of a human being. *Handbook of rational-emotive therapy, 1,* 99-112.

Fani, N., Tone, E. B., Phifer, J., Norrholm, S. D., Bradley, B., Ressler, K. J., ... & Jovanovic, T. (2012). Attention bias toward threat is associated with exaggerated fear expression and impaired extinction in PTSD. *Psychological medicine, 42*(03), 533-543.

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., ... & Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLOS med, 10*(11), e1001547.

Foa, E. B., Feske, U., Murdock, T. B., Kozak, M. J., & McCarthy, P. R. (1991). Processing of threat-related information in rape victims. *Journal of abnormal psychology, 100*(2), 156-62.

Franck, E., De Raedt, R., & De Houwer, J. (2007). Implicit but not explicit self-esteem predict future depressive symptomatology. *Behaviour research and therapy, 45*(10), 2448-55.

Franck, E., De Raedt, R., & De Houwer, J. (2008). Activation of latent self-schemas as a cognitive vulnerability factor for depression: The potential role of implicit self-esteem. *Cognition and emotion, 22*(8), 1588-1599.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. University of Michigan.

Gemar, M. C., Segal, Z. V., Sagrati, S., & Kennedy, S. J. (2001). Mood-induced changes on the Implicit Association Test in recovered depressed patients. Journal of *Abnormal psychology, 110*(2), 282.

Glashouwer, K.A., & de Jong, P.J, (2010). Disorder-specific automatic self-associations in depression and anxiety: results of the Netherlands study of depression and anxiety. *Psychological medicine, 40*, 1101-11.

Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological bulletin, 77*(5), 340.

Gotlib, I. H., Krasnoperova, E., Yue, D. N., & Joormann, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of abnormal psychology, 113*(1), 127.

Greenwald, A. G., Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of personality and social psychology, 79*(6), 1022–1038.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology, 74(*6), 1464.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., & Banaji, M.R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. Journal of Personality and Social Psychology, 97, 17–41.

Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of personality and social psychology, 60*(6), 895.

Hermans, D., Crombez, G., Vansteenwegen, D., Baeyens, F., & Eelen, P. (2002). Expectancy-learning and evaluative learning in human classical conditioning: Differential effects of extinction. In S. P. Shohov, Advances In Psychology Research (pp. 17-40). Hauppauge, NY: Nova Science Publishers.

Hollon, S. D., DeRubeis, R. J., Shelton, R. C., Amsterdam, J. D., Salomon, R. M., O'Reardon, J. P., ... & Gallop, R. (2005). Prevention of relapse following cognitive therapy vs medications in moderate to severe depression. *Archives of general psychiatry, 62*(4), 417-422.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics, 6*(2), 65-70.

Howell, J. L., Gaither, S. E., & Ratliff, K. A. (2014). Caught in the Middle Defensive Responses to IAT Feedback Among Whites, Blacks, and Biracial Black/Whites. Social *Psychological and Personality Science*, 1948550614561127.

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological bulletin, 140*(2), 339.

Insel, T. (2011). Antidepressants: A complicated picture. Director's Blog, National Institute of Mental Health. Accessed 18 April 2016. <http://www.nimh.nih.gov/about/director/2011/antidepressants-a-complicated-picture.shtml>.

Joiner, T. E., Alfano, M. S., & Metalsky, G. I. (1992). When depression breeds contempt: Reassurance seeking, self-esteem, and rejection of depressed college students by their roommates. *Journal of abnormal psychology, 101*(1), 165.

Kagan, J. (2007). A trio of concerns. *Perspectives on psychological science, 2*(4), 361-376.

Kahneman, D., & Riis, J. (2005). Living, and thinking about it: Two perspectives on life. The Science of Well-Being, 285-304.

Kaiser, L. (1989). Adjusting for baseline: change or percentage change?. *Statistics in medicine, 8*(10), 1183-1190.

Kaplan, M. L., Asnis, G. M., Sanderson, W. C., Keswani, L., De Lecuona, J. M., & Joseph, S. (1994). Suicide assessment: Clinical interview vs. self-report. *Journal of clinical psychology, 50*(2), 294-298.

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Chatterji, S., Lee, S., Ormel, J., ... & Wang, P. S. (2009). The global burden of mental disorders: an update from the WHO World Mental Health (WMH) surveys. Epidemiologia e psichiatria sociale, 18(01), 23-33.

Koch, M. (1999). The neurobiology of startle. Prog Neurobiol, 59(2), 107-28.

LeDoux, J. (1996). Emotional networks and motor control: a fearful view. *Progress in brain research*, 107, 437.

Lemmens, L. H., Roefs, A., Arntz, A., van Teeseling, H. C., Peeters, F., & Huibers, M. J. (2014). The value of an implicit self-associative measure specific to core beliefs of depression. *Journal of behavior therapy and experimental psychiatry, 45*(1), 196-202.

Linehan, M. M., & Nielsen, S. L. (1981). Assessment of suicide ideation and parasuicide: Hopelessness and social desirability. *Journal of consulting and clinical psychology, 49*(5), 773.

Ling, Z., & Han, K. (2009, June 10). How to Analyze Change from Baseline: Absolute or Percentage Change? Retrieved May 20, 2016, from http://www.statistics.du.se/essays/D09_Zhang Ling & Han Kun.pdf

Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., & Blair, R. J. R. (2006). The neural basis of implicit moral attitude—an IAT study using event-related fMRI. *Neuroimage, 30*(4), 1449-1457.

Mann, J., & Currier, D. (2008). Suicide and attempted suicide. In Fatemi, S.H., & Clayton, P.J. (Eds.). (2008). The medical basis of psychiatry (pp. 561-576). Totowa, NJ US: Humana Press. doi:10.1007/978-1-59745-252-6_33

Maren, S., & Quirk, G. J. (2004). Neuronal signalling of fear memory. *Nature reviews neuroscience, 5*(11), 844-852.

McCabe, S. B., & Gotlib, I. H. (1993). Attentional processing in clinically depressed subjects: A longitudinal investigation. *Cognitive therapy and research, 17*(4), 359-377.

Minear, M. & Park, D.C. (2004). A lifespan database of adult facial stimuli. *Behavior research methods, instruments, & computer, 36*, 630-633.

National Alliance of Mental Illness (NAMI) Policy Research Institute (2014). The impact and cost of mental illness: The case of depression. Accessed on 15 Sept. 2014, <http://www.nami.org/Template.cfm?Section=Policymakers_Toolkit&Template=/ContentManagement/ContentDisplay.cfm&ContentID=19043>.

National Center for Health Statistics, Centers for Disease Control, & Prevention (Eds.). (2015). Health, United States, 2013, with special feature on prescription drugs. Government Printing Office.

Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and identity, 2*(3), 223-250.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology, 2*(2), 175.

Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J., & Banaji, M. R. (2010). Measuring the suicidal mind implicit cognition predicts suicidal behavior. *Psychological science, 21*(4), 511-7.

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of abnormal psychology, 109*(3), 504.

Owen, A.M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human brain mapping, 25*(1), 46-59.

Partala, T., & Surakka, V. (2003). Pupil size variation as an indication of affective processing. International journal of human-computer studies, 59(1), 185-198.

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: role of the amygdala and vmPFC. *Neuron, 43*(6), 897-905.

Quirk, G. J., Garcia, R., & González-Lima, F. (2006). Prefrontal mechanisms in extinction of conditioned fear. *Biological psychiatry, 60*(4), 337-343.

Radloff, L. S. (1977). The CES-D scale a self-report depression scale for research in the general population. *Applied psychological measurement, 1*(3), 385-401.

Reinhard, G., Lachnit, H., & König, S. (2006). Tracking stimulus processing in Pavlovian pupillary conditioning. *Psychophysiology, 43*, 73-83.

Risch, A.K., Buba, A., Birk, U., Morina, N., Steffens, M.C., & Stangier, U. (2010). Implicit self-esteem in recurrently depressed patients. *Journal of behavior therapy and experimental psychology, 41*, 199-206.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and social psychology bulletin, 27*(2), 151-161.

Robinson, M. D., & Clore, G. L. (2002). Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychological bulletin, 128*(6), 934.

Rogers, C. R., & Dorfman, E. (1951). *Client-centered: Its Current Practice, Implications, and Theory*. ICON Group International.

Schacter, D.L. (2002). The seven sins of memory: How the mind forgets and remembers. Houghton Mifflin Harcourt.

Shiffman, S., Stone, A.A., & Hufford, M.R. (2008). Ecological momentary assessment. *Annual review of clinical psychology, 4*, 1-32.

Singer J.A., & Salovey P. (1993). The Remembered Self: Emotion and Memory in Personality. New York, NY: The Free Press.

Sobell, L. C., Maisto, S. A., Sobell, M. B., & Cooper, A. M. (1979). Reliability of alcohol abusers' self-reports of drinking behavior. *Behaviour research and therapy, 17*(2), 157-160.

Sobell, L. C., & Sobell, M. B. (1992). Timeline follow-back. In Measuring alcohol consumption (pp. 41-72). Humana Press.

Sowislo, J. F., & Orth, U. (2013). Does low self-esteem predict depression and anxiety? A meta-analysis of longitudinal studies. *Psychological bulletin, 139*(1), 213.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). Manual for the state-trait anxiety inventory. University of Buffalo Institutional Repository. Accessed 7 Oct. 2014. <https://ubir.buffalo.edu/xmlui/handle/10477/2895>.

Stone, A. A., Schwartz, J. E., Neale, J. M., Shiffman, S., Marco, C. A., Hickcox, M., ... & Cruise, L. J. (1998). A comparison of coping assessed by ecological momentary assessment and retrospective recall. *Journal of personality and social psychology, 74*(6), 1670.

Thomaz, C. E., & Giraldi. G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and vision computing, 28*(6), 902-913.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-31.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology, 54*(6), 1063.

Wenze S.J., Gunthert K.C., & German R.E (2012). Biases in affective forecasting and recall in individuals with depression and anxiety symptoms. *Personality and social psychology bulletin, 38*(7): 895-906.

Wilson T.D. (2009). Know thyself. *Perspectives on psychological science, 4*, 384-9.

Yoshihama, M., Clum, K., Crampton, A., & Gillespie, B. (2002). Measuring the lifetime experience of domestic violence: application of the life history calendar method. *Violence and victims, 17*(3), 297-317.

## Figures

Figure 1. Neural circuit of fear conditioning and extinction

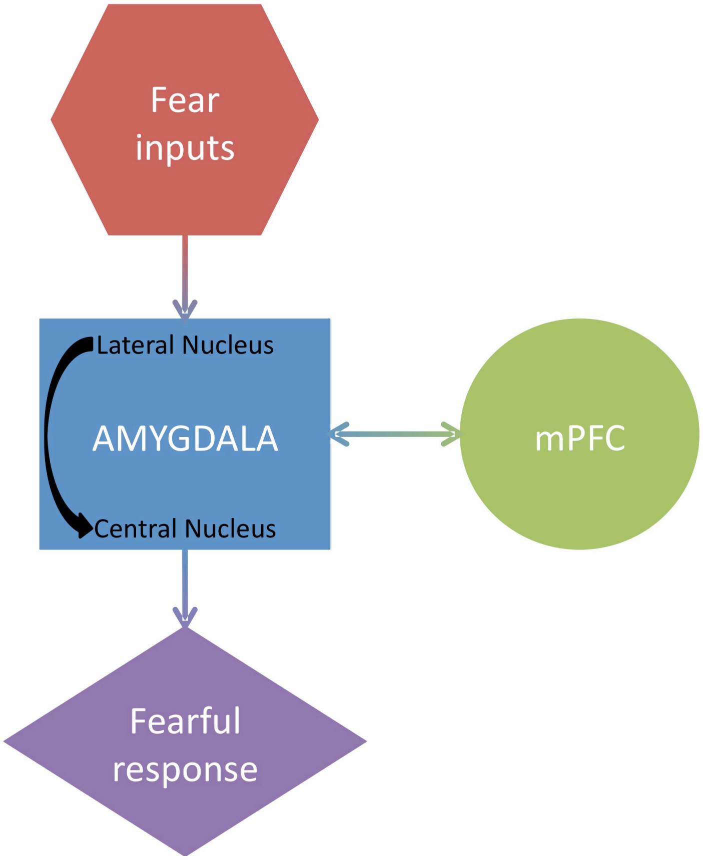


Figure 2. Experimental design of Study 1

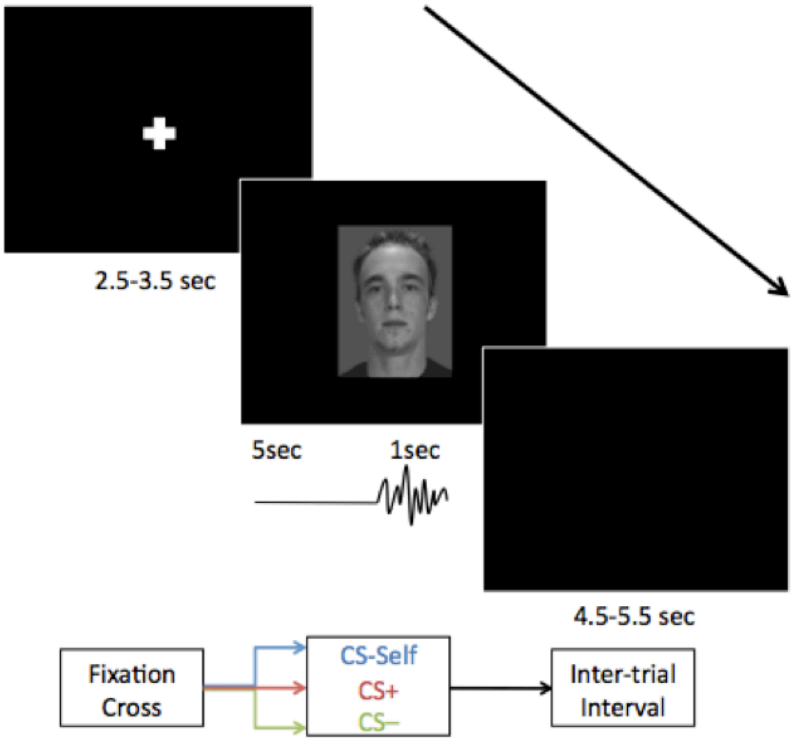


Figure 3. Experimental design of Study 2

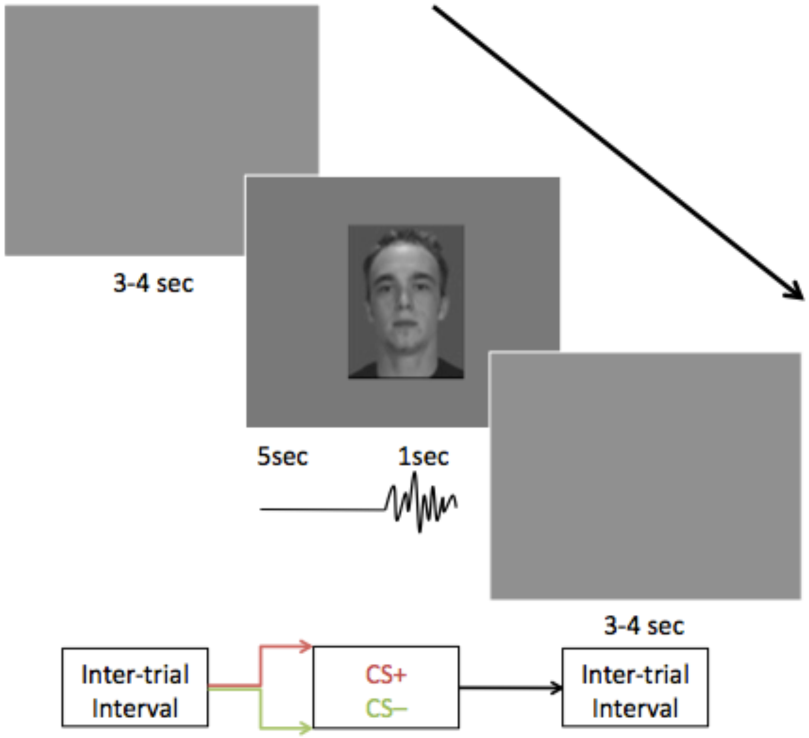


Figure 4. Experimental design of Study 3

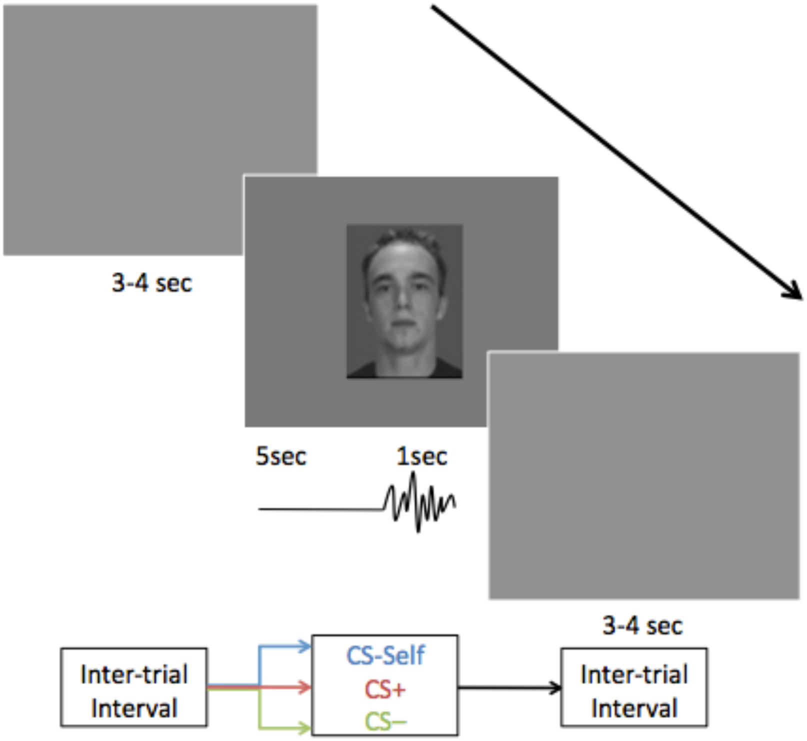


Figure 5. Study 1: Graph of percent change from average pupil area of fixations falling in interest areas during "clean" time from all trials in that block.
ACQ1 = first block of acquisition; ACQ2 = second block of acquisition; EXT1 = first block of extinction 1; EXT2 = second block of extinction. Error bars represent +/– 1SE

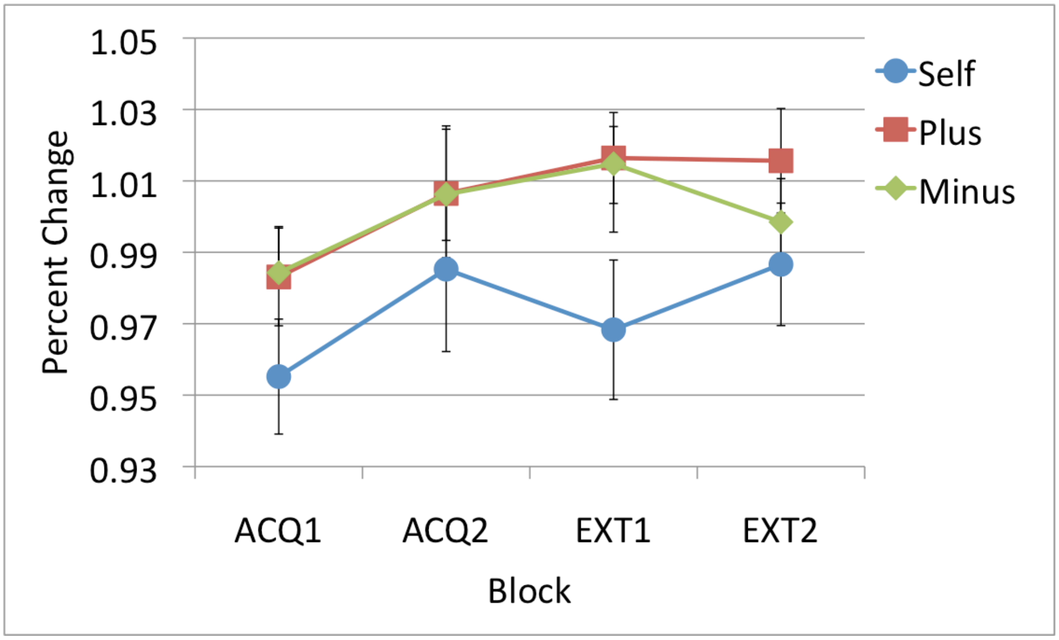


Figure 6. Study 2: Graph of percent change from average pupil area of fixations falling in interest areas during "clean" time from all trials in that block
ACQ1 = first block of acquisition; ACQ2 = second block of acquisition; EXT1 = first block of extinction 1; EXT2 = second block of extinction. Error bars represent +/– 1SE

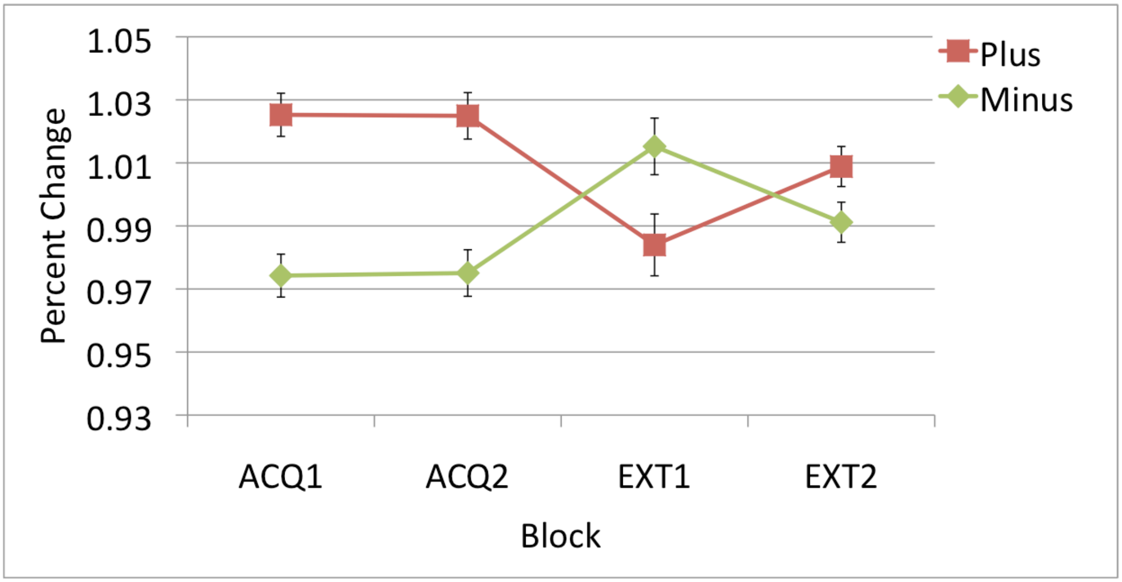


Figure 7. Study 3: Graph of percent change from average pupil area of fixations falling in interest areas during "clean" time from all trials in that block
ACQ1 = first block of acquisition; ACQ2 = second block of acquisition; EXT1 = first block of extinction 1; EXT2 = second block of extinction. Error bars represent +/– 1SE

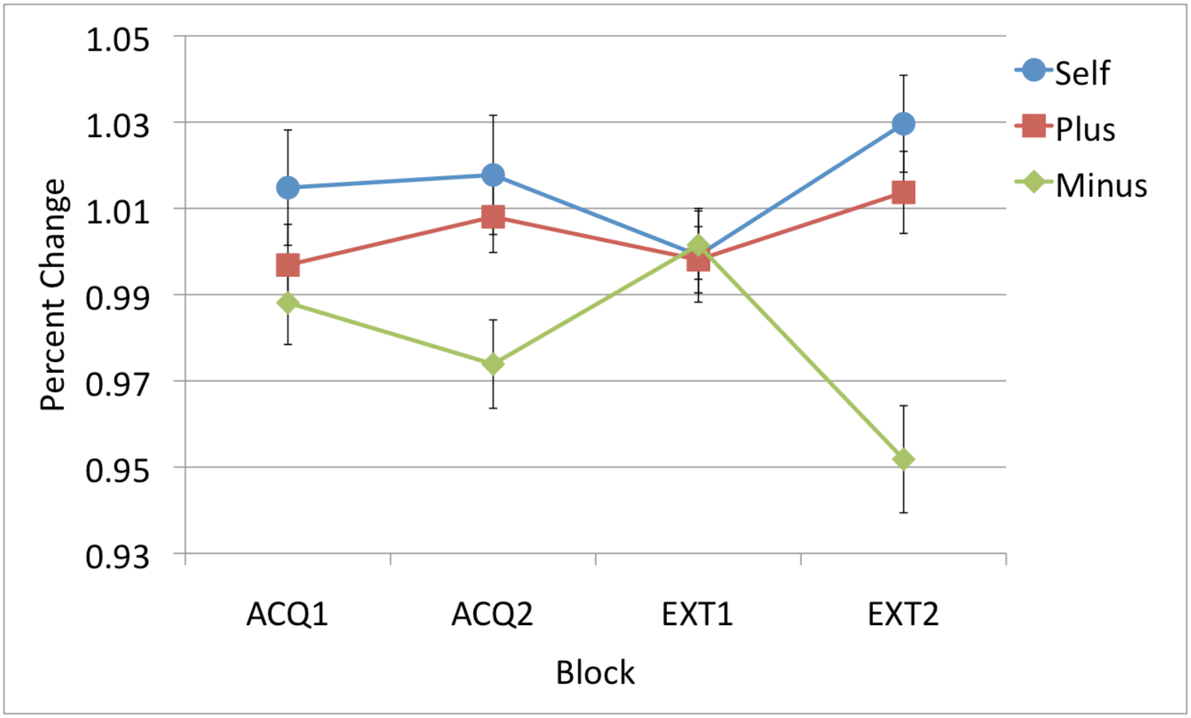
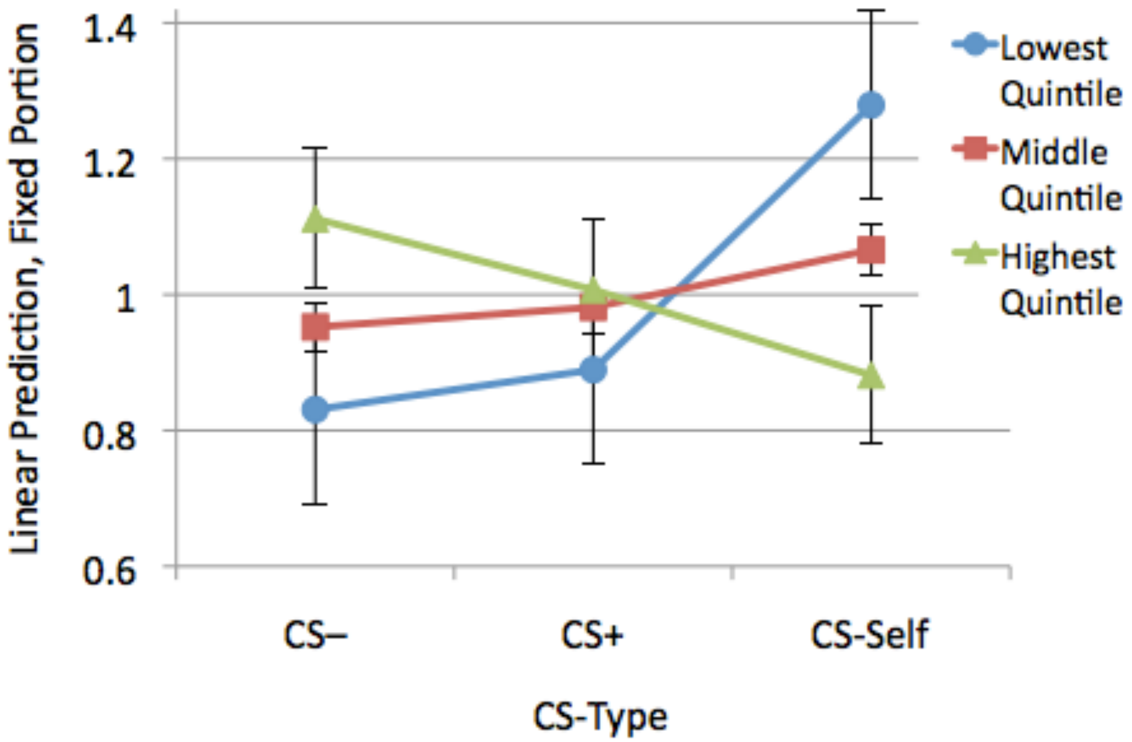


Figure 8. Graph of marginal means from condition*NSCS in acquisition for highest, middle and lowest quintile NSCS scores.


Linear Prediction, Fixed Portion
CS–
CS+
CS-Self
CS-Type
Lowest Quintile
Middle Quintile
Highest Quintile