**Article:**

Izuma, Keise orcid.org/0000-0003-0256-3571, Kennedy, Kate, Fitzjohn, Alexander et al. (2 more authors) (2018) Neural activity in the reward-related brain regions predicts implicit self-esteem: A novel validity test of psychological measures using neuroimaging. Journal of Personality and Social Psychology. 343–357. ISSN 0022-3514

https://doi.org/10.1037/pspa0000114

**DOI: 10.1037/pspa0000114**

# Neural Activity in the Reward-Related Brain Regions Predicts Implicit Self-Esteem: A Novel Validity Test of Psychological Measures Using Neuroimaging

Keise Izuma, Kate Kennedy, and Alexander Fitzjohn

University of York, UK

Constantine Sedikides

University of Southampton, UK

Kazuhisa Shibata

Nagoya University, Japan

Corresponding author: Keise Izuma, Department of Psychology, University of York, Heslington, York, YO10 5DD, UK; Tel: +44 (0)1904 323167; Email: keise.izuma@york.ac.uk

## Abstract

Self-esteem, arguably the most important attitudes an individual possesses, has been a premier research topic in psychology for more than a century. Following a surge of interest in implicit attitude measures in the 90s, researchers have tried to assess self-esteem implicitly in order to circumvent the influence of biases inherent in explicit measures. However, the validity of implicit self-esteem measures remains elusive. Critical tests are often inconclusive, as the validity of such measures is examined in the backdrop of imperfect behavioral measures. To overcome this serious limitation, we tested the neural validity of the most widely used implicit self-esteem measure, the implicit association test (IAT). Given (1) the conceptualization of self-esteem as attitude toward the self, and (2) neuroscience findings that the reward-related brain regions represent an individual's attitude or preference for an object when viewing its image, individual differences in implicit self-esteem should be associated with neural signals in the reward-related regions during passive-viewing of self-face (the most obvious representation of the self). Using multi-voxel pattern analyses (MVPA) on functional magnetic resonance imaging (fMRI) data, we demonstrated that the neural signals in the reward-related regions were robustly associated with implicit (but not explicit) self-esteem, thus providing unique evidence for the neural validity of the self-esteem IAT. In addition, both implicit and explicit self-esteem were related, although differently, to neural signals in regions involved in self-processing. Our finding highlights the utility of neuroscience methods in addressing fundamental psychological questions and providing unique insights into important psychological constructs.


Keywords: self-esteem, fMRI, MVPA, IAT, implicit attitude, implicit measure

## Neural Activity in the Reward-Related Brain Regions Predicts Implicit Self-Esteem: A Novel Validity Test of Psychological Measures Using Neuroimaging

In the past two decades, implicit attitude measures (most prominently, the Implicit Association Test [IAT]; Greenwald, McGhee, & Schwartz, 1998) have attracted a surge of interest from scientists and the public as a tool to uncover unconscious attitudes, that is, attitudes that an individual is unable or unwilling to report. Still, some remain skeptical of implicit measures' validity (Blanton, Jaccard, Christie, & Gonzales, 2007; Blanton et al., 2009). Among all attitude domains to which implicit measures have been applied, no domain has attracted more skepticism than self-esteem. Implicit methods to measure self-esteem have been criticized as lacking sufficient validity (i.e., low convergent and predictive validity, low test-retest reliability) (Bar-Anan & Nosek, 2014; Bosson, Swann, & Pennebaker, 2000; Buhrmester, Blanton, & Swann, 2011; Falk & Heine, 2015; Falk, Heine, Takemura, Zhang, & Hsu, 2015; Rudolph, Schroder-Abe, Schutz, Gregg, & Sedikides, 2008), and some authors have even concluded in favor of invalidity (Buhrmester et al., 2011; Falk et al., 2015).

It is difficult, however, to make a definitive contribution to that debate, because validity has been assessed in reference to other imperfect behavioral measures. For example, Falk et al. (2015) collected nine implicit measures of self-esteem from three groups of participants (Euro-Canadians, Asian-Canadians, Japanese). The implicit measures were uncorrelated with each other across all three groups, demonstrating the low convergent validity of implicit self-esteem measures. However, we cannot conclude from these results that all implicit self-esteem measures are invalid: even if one measure was perfectly reliable and valid, no correlation would emerge in the case in which all other measures were invalid.

Similarly, the low predictive validity of implicit self-esteem measures found in prior research may be due to biases in selecting criterion variables. Researchers have typically selected criterion variables based on understanding of what explicit self-esteem is (Bosson et al., 2000; Falk et al., 2015). As a consequence, almost all criterion variables have been strongly correlated with explicit self-esteem measures, but not with implicit self-esteem measures (Bosson et al., 2000; Falk et al., 2015; for a review, see Buhrmester et al., 2011). Given the divergent validity of implicit and explicit self-esteem (Bosson et al., 2000; Buhrmester et al., 2011; Falk et al., 2015; Greenwald & Farnham, 2000; Rudolph et al., 2008), this literature may not be a fair test of the predictive validity of implicit self-esteem measures. Stated otherwise, lack of predictive validity may simply reflect unclarities in the definition of implicit self-esteem.

We aim to overcome this methodological and conceptual limitation and provide independent evidence for the validity of an implicit self-esteem measure. In particular, we investigate whether implicit self-esteem, as measured by the IAT, is associated with robust neural representations. We focused on the IAT, because it is more reliable than other implicit measures in terms of internal consistency and test-retest reliability (Bosson et al., 2000; Krause, Back, Egloff, & Schmukle, 2011; Rudolph et al., 2008). We emphasize that, although we use a neuroimaging method, our primary objective is to address a psychological question (i.e., the validity of an implicit self-esteem measure) rather than a neuroscience question (e.g., neural correlates of implicit self-esteem). We thus adopt a neuroimaging approach known as *psychological hypothesis testing* (Amodio, 2010).

More specifically, we test whether self-esteem IAT scores are robustly associated with neural activation in the reward-related brain regions (Bartra, McGuire, & Kable, 2013; Kolling, Behrens, Wittmann, & Rushworth, 2016; Schultz, 2015; Sescousse, Caldu, Segura, & Dreher, 2013) in response to self-face—arguably, the most obvious, immediate, and authentic representation of the self. Previous neuroimaging studies demonstrated that incidental preferences or attitudes toward various stimuli are automatically represented (i.e., without asking participants to report how they feel about the stimuli) in the reward-related areas, such as striatum and ventromedial prefrontal cortex (vmPFC) (Izuma, Shibata, Matsumoto, & Adolphs, 2017; Lebreton, Jorge, Michel, Thirion, & Pessiglione, 2009; Levy, Lazzaro, Rutledge, & Glimcher, 2011; Smith, Bernheim, Camerer, & Rangel, 2014; Tusche, Bode, & Haynes, 2010), and that individual differences in neural activities in these regions in response to rewarding stimuli are correlated with self-reported positive affect or preference for the stimuli (Bjork et al., 2004; Hariri et al., 2006; Knutson, Adams, Fong, & Hommer, 2001; Knutson, Taylor, Kaufman, Peterson, & Glover, 2005; Wu, Bossaerts, & Knutson, 2011). Furthermore, prior neuroimaging studies have shown the involvement of these reward related regions in explicit (but not implicit) self-esteem, as measured by a standardized questionnaire (i.e., trait self-esteem) (Chavez & Heatherton, 2015; Frewen, Lundberg, Brimson-Theberge, & Theberge, 2013; Oikawa et al., 2012) as well as momentary shift in how individuals feel about themselves (i.e., state self-esteem; Will, Rutledge, Moutoussis, & Dolan, 2017). The results of a more recent study (Chavez, Heatherton, & Wagner, 2017) also indicated that people's tendency to view themselves in a positive manner is reflected in neural activations in the vmPFC, suggesting that, like preferences for consumer goods, positive attitudes toward the self are associated with activity in reward-related brain regions. In other words, neural responses in the reward-related brain regions while viewing self-face is an appropriate criterion variable, because of a close theoretical fit between what the self-esteem IAT scores and the neural responses should reflect (i.e., automatic evaluation of the self).

Thus, given that self-esteem is often conceptualized as attitude toward the self (Sedikides

& Gregg, 2003), and implicit self-esteem is commonly defined as the association of the concept of self with positive or negative valence (Greenwald et al., 2002), if the IAT is a valid measure of self-esteem, its scores should be associated with neural signals in the reward-related brain regions. Stated otherwise, if self-esteem IAT scores did not reflect individual differences in any meaningful psychological trait (Buhrmester et al., 2011; Falk et al., 2015), it would be highly unlikely to observe an association between self-esteem IAT scores and neural signals in the reward-related brain regions.

In doing so, we employed a functional neuroimaging technique (functional magnetic resonance imaging or fMRI) combined with a machine learning technique called multi-voxel pattern analysis (MVPA; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006). MVPA is known to be more sensitive in detecting different psychological, cognitive, or perceptual states than conventional fMRI data analysis (Izuma et al., 2017; Jimura & Poldrack, 2012; Sapountzis, Schluppeck, Bowtell, & Peirce, 2010) and thus suitable for identifying potentially complex associations between implicit self-esteem and neural signals in reward-related brain regions (see Methods for more details). Indeed, using MVPA, a previous fMRI study (Ahn et al., 2014) demonstrated that it is possible to predict individual differences in attitudes (political ideology) based on brain activities. Ahn et al. (2014) found that a correlation between actual political attitudes measured by a questionnaire and predicted attitudes based on MVPA was fairly high ($r$ = 0.82), suggesting that MVPA is a reliable method for identifying the relation between an attitude measure and brain activities.

We scanned the brains of 43 individuals via fMRI while presenting them with pictures of their own face (Figure 1; see Methods for power analysis). We instructed participants to carry out a simple attention task while viewing pictures; we did not ask them to consider how they felt about themselves. Following the fMRI scanning, each participant completed the self-esteem IAT (Greenwald & Farnham, 2000) as well as two explicit self-esteem measures: (1) Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) and (2) State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991). By applying MVPA to the fMRI data, we were able to test whether participants' level of implicit self-esteem was reliably predicted from neural signals obtained while viewing their own faces. We further examined whether explicit self-esteem scores (RSES) can be similarly predicted by neural signals in the reward-related brain regions, aiming to provide evidence for the divergent validity of implicit versus explicit self-esteem.

## Method

### Participants

We recruited 48 women from the Neuroimaging Centre participant pool. All participants were current students at the University of XXX. The literature suggests gender differences in self-esteem (Bleidorn et al., 2016; Kling, Hyde, Showers, & Buswell, 1999) as well as in the relationship between perceived self-face attractiveness and self-esteem (Pliner, Chaiken, & Flett, 1990). Thus, while passive viewing of self-face would induce neural signals related to automatic evaluation of the self in both genders, the sensitivity of such responses might differ across genders. Accordingly, we recruited only females in an effort to bypass such differences in this first, validation study. Other inclusion criteria were: (1) ages of 18 to 28, (2) right-handedness[1], (3) native command of the English language, (4) no history of neurological or psychiatric illness, and (5) no metal body implants or devices. We excluded five participants from the analyses: Three of them did not complete the fMRI session (two due to a problem with an fMRI scanner,

[1] The literature has pointed to differences in brain anatomy between right-handers and left-handers (e.g., Toga & Thompson, 2003). Thus, following a typical procedure of neuroimaging studies, we limited our sample to right-handed individuals.

one due to her decision to withdraw), and the remaining two were identified to have a brain anomaly. The final sample consisted of 43 participants aged 18-28 years (*M* = 20.9, *SD* = 2.46). All participants provided written informed consent. Ethics approval was granted by the Ethics Board of University of XX.

**Power Analysis**

We estimated the effect size to be $r = 0.392$ based on a previous investigation (Ahn et al., 2014). As in the present study, Ahn et al. (2014) attempted to predict individual difference in social attitudes on the basis of fMRI signals. They focused on political attitudes, and reported that the correlation between predicted and actual attitudes across participants ($N = 83$) was $r = 0.82$. One crucial difference between Ahn et al.'s investigation and the present study is that our behavioral measure (i.e., IAT) is likely to be noisier than their measure of political attitudes. We estimated the difference in measurement noise based on test-retest reliability. Ahn et al. (2014) reported that the test-retest reliability of political attitudes was $r = 0.952$, whereas the test-retest reliability of the self-esteem IAT is $r = 0.455$; this is the average reliability of the following five studies (weighted by number of participants): $r = 0.69$ (Bosson et al., 2000), $r = 0.54$ (Krause et al., 2011), $r = 0.54$ (Rudolph et al., 2008, Study 1), $r = 0.52$ (Greenwald & Farnham, 2000), $r = 0.39$ (Rudolph et al., 2008, Study 3), and $r = 0.31$ (Gregg & Sedikides, 2010). Based on this information, we estimated an effect size of $r = 0.392$ for our study. With such an effect size, a sample size of $n = 39$ would achieve statistical power of $\beta = 0.2$, $\alpha = 0.05$ (one-tailed). In order to account for potential data loss (e.g., due to excessive head motion in the scanner), we aimed to recruit a total of 45 participants and ended up recruiting 48.

**Pre-Screening**

To ensure that our sample was characterized by a wide range of self-esteem, we emailed those who expressed an interest in our fMRI study, asking them to complete an online questionnaire which included the RSES. A total of 167 individuals completed the questionnaire. 129 of the 167 respondents were eligible for the fMRI experiment (e.g., female, 18-29 years-old, right-handed, native English speakers, no history of neurological or psychiatric illness, no metal in the body). The self-esteem scores of these 129 respondents were normally distributed (*range* = 8-30, $M$ = 19.14, $SD$ = 4.66). We invited them all for participation in the fMRI study, except for most of those whose self-esteem scores hovered around the mean (16-24). Of note, the self-esteem statistics (RSES score) for our final sample ($n$ = 43) at the pre-screening stage were: *range* = 8-30, $M$ = 19.88, $SD$ = 5.39.

**Stimuli**

We employed images of participants' own faces as experimental stimuli during the fMRI scanning (Figure 1a). For use in the self-image presentation inside an fMRI scanner, we took four photographs of each participant under uniform lighting conditions during a 15-minute session a few weeks prior to scanning with a Nikon Coolpix s9900 digital camera (1600 × 1200 pixels). Photographs were front facing passport style, with participants displaying neutral, open-eyed expressions. We also used four scrambled images of natural scenes (i.e., not self-images; Figure 1b) as emotionally-neutral control stimuli, so that all participants viewed the same scrambled images.

------------------ Insert Figure 1 about here -------------------

We selected scrambled images as control stimuli, because we considered them emotionally neutral. Given that we aimed to predict *individual differences* in self-esteem from neural signals, an ideal control stimulus would induce the same attitude-related activations across all participants (e.g., neutral for everyone). It could be argued that control stimuli like faces of unfamiliar individuals are more appropriate, as they have been used in prior research (Kaplan, Aziz-Zadeh, Uddin, & Iacoboni, 2008; Sugiura et al., 2000). However, this research was concerned with brain regions specific to self-faces, and thus its objective was fundamentally different from the objective of the present study. Faces of unfamiliar individuals are not suitable control stimuli in our study: There are individual differences in face attractiveness judgement (Honekopp, 2006), and facial attractiveness/trustworthiness affects neural activity in reward-related brain regions (Mende-Siedlecki, Said, & Todorov, 2013). Hence, use of unfamiliar individuals' faces as control stimuli would likely reduce signals in which we were interested.

Furthermore, it could be argued that, because there are so many differences between self-face and scrambled images, we cannot make strong inferences based on contrasts between these conditions. There are two key differences between the present study and typical neuroimaging research. First, again, the present study does not aim to identify brain regions specific to self-face processing. Second, we used a machine learning technique (MVPA; see below for more detail) to detect activation patterns that are associated with individual differences in the automatic evaluation of the self (implicit self-esteem). Machine learning is capable of locating specific patterns that are associated with a variable of interest from big (and noisy) data (Alpaydin, 2014). As stated above, neural signals related to individual differences in the automatic evaluation of the self should be included in the contrast of the self-face versus scrambled image conditions (especially in reward-related brain regions). If so, machine learning (MVPA) should be able to locate specific information related to it and thus predict implicit self-esteem.

**Procedure**

The study consisted of two sessions on two separate days: (1) photo session, and (2) fMRI session. On the first day, we asked participants to complete the photo session. After we gave them general instructions on the project and fMRI safety information, we took four photographs of each participant. The photo session occurred an average of 27 days prior to the fMRI experiment. We concealed the true purpose of the study (i.e., predicting self-esteem based on brain activities) by mentioning to participants that it was concerned with neural responses to social versus non-social objects.

On the second day, during fMRI scanning, participants viewed 30 blocks. These were (1) self-images blocks, (2) scrambled-image control blocks, and (3) rest (i.e., a fixation cross) blocks (10 blocks each). Presentation of each block lasted 12 sec. In each of the self-image and scrambled-image blocks, we presented 4 different images for 2 sec each in randomized order per block (inter-stimulus interval = 1 sec). Within each block, at random intervals one image darkened for 300ms, which participants were instructed to detect and respond to as quickly as possible with a right index finger button press. We asked participants to engage in this simple task inside the scanner in order to ensure that they were paying attention to the presented images. Similar low-demanding tasks have been used in studies that examined neural responses related to automatic evaluations of various stimuli (Ahn et al., 2014; Cunningham et al., 2004; Izuma et al., 2017; Smith et al., 2014). We recorded participants' responses within a 2 sec window post-luminance change. Given that we were interested in how individual differences in implicit self-esteem are related to brain activations, we fixed the order of blocks across all participants. After the fMRI run (a total of 6 min), each participant underwent a different fMRI run, which is unrelated to the objective of the current study (and the relevant data will not be reported here).

Following fMRI scanning, we instructed participants to engage in behavioral tasks.

Participants first completed a self-esteem IAT (Greenwald & Farnham, 2000). We created the IAT with Psychopy stimulus presentation software (Peirce, 2007). The IAT comprised the four following catetories: (1) Self, (2) Other, (3) Positive, and (4) Negative. The Self category included *I, My, Me, Mine,* and *Self*, whereas the Other category included *they, them, their, theirs* and *other*. In addition, the Positive category included 10 positive words (e.g., *Peace, Joy, Honest*), whereas the Negative category included 10 negative words (e.g., *Agony, Stupid, Useless*).

Following the IAT, we administered the explicit self-esteem measures of RSES and SSES. Note that the SSES consists of three sub-scales: appearance, performance, and social. The subscales assess aspects of self-esteem that are based on physical appearance, ability, and others' evaluation, respectively. Finally, participants rated the attractiveness of their face (“how attractive do you think your face is compared to average students on campus”) on a 7-point scale (1 = *Least Attractive*, 4 = *Average*, 7 = *Most Attractive*). Upon completion, we paid participants £16 and debriefed them.

**Behavioral Data Analysis**

We calculated a self-esteem IAT score for each participant using the algorithm developed by Greenwald, Nosek, and Banaji (2003). We excluded one participant from the analyses of the behavioral data obtained during the fMRI scanning (reaction time and performance in the luminance change detection task) due to malfunction of the response box. For paired t tests, following Equation 3 of Dunlap, Cortina, Vaslow, and Burke (1996), we computed the effect sizes by

$$d = t[2(1 - r)/n]^{1/2}$$

where *t* is the t-statistic, *r* is the correlation between two measures, and *n* is the sample size.

### fMRI Data Acquisition

We used an 8 Channel head coil, GE 3T HDx Excite MRI scanner in the Neuroimaging Centre to acquire whole brain fMRI data. Participants underwent a 13 second standard localizer scan and 12 second ASSET calibration for parallel imaging. We also obtained high resolution T1-structural scans (TE = 3 minute minimum full; TR = 7.8ms; TI = 450ms; 20° flip angle matrix = 256x256x176; FOV = 290x290x176; slice thickness = 1.13x1.13x1mm voxel size). Functional data collection consisted of a 6 min scan, gathering 120 volumes using T2*-sensitive echo-planar imaging (TE = 30ms; TR = 3000ms; 90° flip angle; matrix = 96x96; FOV = 288mm). We used horizontal orientation interleaved bottom-up acquisition, with 38 4mm slices (128x128 voxels per slice; 2mm voxel).

### **fMRI Data Pre-processing**

We analyzed the fMRI data using SPM8 (Wellcome Department of Imaging Neuroscience) implemented in MATLAB (MathWorks). Before data processing and statistical analysis, we discarded the first four volumes to allow for T1 equilibration. Following motion correction, we normalized the volumes to MNI space using a transformation matrix obtained from the normalization of the first EPI image of each participant to the EPI template (resliced to a voxel size of 3.0 × 3.0 × 3.0 mm). We used these normalized data for the MVPA data analyses. We spatially smoothed the normalized fMRI data with an isotropic Gaussian kernel of 8 mm (full-width at half-maximum). We used the smoothed fMRI data for MVPA analyses on the basis of research showing that smoothing can improve decoding performance when large-scale activation patterns are assumed (Op de Beeck, 2010).

### **Univariate fMRI Data Analysis**

We first ran a conventional general linear model (GLM) analysis where the signal time course for each participant was modeled with a GLM (Friston et al., 1995). In the GLM, we modeled separately (duration = 12 sec) each of the self and scrambled-image blocks. We generated regressors of interest (condition effects) using a box-car function convolved with a hemodynamic-response function. We excluded regressors that were of no interest: six head motion parameters (translation: x, y, and z; rotations: pitch, roll, and yaw) and high-pass filtering (128 s). We created a contrast image for Self-image versus Scrambled-image for each participant, and used it in subsequent MVPA analyses (see below).

Furthermore, in the second level analysis, for the Self-image versus Scrambled-image contrast, we entered implicit (IAT) and explicit (RSES) self-esteem scores as covariates to test whether implicit or explicit self-esteem were linearly related to activations in reward-related brain regions. For the univariate analysis, we reasoned that the effect size (i.e., correlation between implicit self-esteem scores and brain activity) should be, if anything, lower than the effect size based on the MVPA mentioned above, due to the lower sensitivity of univariate analysis. Accordingly, for the reward-related regions (see below for more detail on how we defined a region of interest [ROI]), we used a slightly lenient statistical threshold of $p < 0.01$ voxelwise (uncorrected; note that $p = 0.01$ corresponds to $r = 0.354$) and cluster $p < 0.05$ (FWE corrected for multiple comparisons). For the regions outside of the reward related ROI, we set the statistical threshold at $p < 0.005$ voxelwise (uncorrected) and cluster $p < 0.05$ (FWE corrected for multiple comparisons).

**MVPA**

In order to predict self-esteem IAT scores from neural signals, we employed MVPA (Haynes & Rees, 2006; Norman et al., 2006). In contrast to the traditional fMRI data analysis approach that only evaluates univariate change in voxel-wise intensity, the MVPA is considered and proven to be more sensitive in detecting and distinguishing cognitive states in the brain (e.g., Izuma et al., 2017; Jimura & Poldrack, 2012; Sapountzis et al., 2010), because it allows researchers to extract the signal that is present in the pattern of brain activations across multiple voxels. For example, with the conventional univariate analysis, we could identify the relation between self-esteem and neural activity only if the strength of activation was positively (or negatively) related to individuals' self-esteem scores (e.g., the higher the activation in an area, the higher the self-esteem scores). In contrast, even if there is no difference in overall activation strength across individuals with different level of self-esteem, there may be specific differences in activation *patterns* across multiple voxels, and, if so, a machine learning algorithm could identify the patterns that explain (predict) self-esteem scores.

We used in particular a machine learning algorithm called support vector regression (SVR; Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) as implemented in LIBSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/), with a linear kernel and a cost parameter of c = 1 (default). We also set all other parameters to their default values. We previously used the SVR and successfully predicted participants’ attitudes toward familiar celebrities from brain activations obtained during passive-viewing of these celebrities (reference omitted for masked review purposes).

We computed prediction performance using the 6-fold balanced cross-validation procedure (Cohen et al., 2010; see also Kohavi, 1995); we first divided participants into 6 groups (7-8 participants in each group), such that these 6 groups had roughly the same means and variances of self-esteem IAT scores (or RSES scores when predicting explicit self-esteem). We left out one group in each cross-validation and conducted the SVR using the data from participants in all other groups (training data). The SVR uses the training data to learn a weight value for each voxel in a ROI, which represents the contribution of a particular voxel to predicting self-esteem scores. Then, these weights are tested on participants in the left-out group (predicted IAT scores for each participant in the left-outgroup is computed based on their neural signals). We repeated this procedure for each group (a total of 6 times), and computed a Pearson's correlation coefficient between actual IAT scores and predicted scores.

We tested whether brain activations in the reward-related regions predicted self-esteem IAT scores. We defined the reward-related brain areas based on Neurosynth (http://www.neurosynth.org/; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). We used an activation map from the term "Reward" (reverse inference map only), thresholded at $q < 0.01$ False Discovery Rate (FDR) corrected. This ROI (Figure 2a; a total of 2,696 voxels; note that we used the largest cluster only) comprises brain regions that are preferentially implicated in neuroimaging studies, which addressed the neural bases of reward processing[2] and included areas involved in reward processing such as vmPFC, caudate nucleus, and midbrain (Figure 2a). We also conducted the same analysis using a ROI defined by a meta-analysis (Bartra et al., 2013). This meta-analysis identified a network of brain regions positively associated with subjective value including bilateral striatum, vmPFC, bilateral insula, anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), and midbrain (brainstem). This amount to a total of 3,680 voxels; see Figure 3A in Bartra et al., 2013).

To check the robustness of the results obtained with the reward ROI (Figure 2a), we also

[2] More precisely, in the term ("Reward") based meta-analysis, Neurosynth employs text-mining techniques to identify neuroimaging studies that used the term "Reward" at a high frequency, extract activation coordinates reported in all tables, and run meta-analyses (Yarkoni et al., 2011). Therefore, it is possible that not all studies included in the meta-analysis addressed the neural bases of reward processing.

ran MVPA using the following two ROIs. *First*, the large reward ROI (Figure 2a) included medial prefrontal cortex (mPFC) regions, especially its ventral part (vmPFC). Given that mPFC is known to be involved in self-processing (Denny, Kober, Wager, & Ochsner, 2012; Northoff et al., 2006), which might be related to implicit or explicit self-esteem, we excluded these regions from the reward ROI by applying anatomical masks (in particular, vmPFC, mPFC, and anterior cingulate cortex [ACC]) using a WFU pickatlas toolbox for SPM (Maldjian, Laurienti, Kraft, & Burdette, 2003). The new ROI (Figure 3a) consists of a total of 2,179 voxels. *Second*, in order to limit our ROI only to regions that are even more selective to reward, we thresholded the reverse-inference map obtained from Neurosynth (Figure 2a) at z-score = 10.[3] The higher threshold eliminated not only regions in the frontal cortex (e.g., vmPFC, ACC) but also other regions (e.g., putamen, thalamus, amygdala) that are relatively less selective to reward. The new ROI (Figure 3b) consists only of bilateral ventral striatum (nucleus accumbens) and midbrain (a total of 343 voxels), which are known to be the center of the reward circuit (Haber & Knutson, 2010). It is well established that midbrain is rich in dopamine neurons, which encode reward-related information (e.g., reward prediction error; Schultz, 2015). Similarly, ventral striatum (nucleus accumbens), which is heavily interconnected with midbrain (Haber & Knutson, 2010), is known to be highly sensitive (Bartra et al., 2013; Sescousse et al., 2013) and is selective to reward (Ariely & Berns, 2010).

To examine further if each anatomical region in the reward-related brain regions accounts for individual difference in self-esteem, we selected 13 reward-related anatomical structures based on the abovementioned reverse inference map from Neurosynth (Figure 2a): (1) vmPFC; (2) left caudate nucleus; (3) right caudate nucleus; (4) left pallidum; (5) right pallidum; (6) left

[3] We selected z-score = 10, because with any z-score lower than 10, there were active voxels in the frontal cortex.

putamen; (7) right putamen; (8) ACC; (9) left amygdala; (10) right amygdala; (11) left thalamus; (12) right thalamus; and (13) midbrain. Each of the 13 reward-ROIs are known to contain neurons that encode rewards or values (Komura et al., 2001; Mizuhiki, Richmond, & Shidara, 2012; Nishijo, Ono, & Nishino, 1988; Schultz, Apicella, & Ljungberg, 1993; for reviews, see: Kolling et al., 2016; Schultz, 2015) and has been consistently activated in response to various types of social and non-social rewards in human neuroimaging studies (Bartra et al., 2013; Izuma, 2015; Sescousse et al., 2013). We also examined whether self-esteem scores could be predicted by activation patterns in areas that were not previously implicated in reward. We selected the non-reward related anatomical ROIs as follows. *First*, using Neurosynth, we obtained another activation map from the term "Reward," but this time the map included both reverse and forward inference maps, thresholded at $q < 0.05$ FDR corrected. This map (a total of 5,605 voxels) includes brain regions that were consistently (but not necessarily selectively) activated in previous studies which focused on the neural bases of reward processing. *Second*, we selected all anatomical structures that are not included in this broadly-defined reward-related regions. These non-reward ROIs mainly include areas in parietal, temporal and occipital cortices (a total of 55 non-reward ROIs; see Table 3 below for the full list of the 55 ROIs). We created all of the anatomical ROIs using a WFU pickatlas toolbox for SPM (Maldjian et al., 2003).

Similarly, for exploratory MVPA analyses, we defined self-related brain regions using Neurosynth. We used an activation map from the term "Self" (reverse inference map only), thresholded at $q < 0.01$ FDR corrected. This ROI consists of two cluster (Figure 5a): (1) mPFC (421 voxels), and (2) posterior cingulate cortex (PCC; 186 voxels), both of which are areas previously identified in meta-analyses of fMRI studies on self-processing (Denny et al., 2012; Northoff et al., 2006).

We evaluated prediction performance in each ROI with a permutation test (Shibata, Watanabe, Kawato, & Sasaki, 2016). We created 5,000 randomly shuffled permutations of self-esteem scores (both IAT and RSES; note that we shuffled the scores within each of the 10 fold groups so that the averages scores in the 10 fold groups were maintained across the permutations) and ran the SVR using the permutated data in each ROI to obtain a distribution of correlations between predicted and actual self-esteem under the null hypothesis. Thus, this distribution tells us how unlikely it is to obtain the results we found, if the self-esteem IAT score reflected noise. After the MVPA analyses, correlation coefficients between actual self-esteem scores and predicted scores were Fisher-z transformed before any further analysis. Notably, as RSES scores were highly correlated with a total SSES scores as well as each of 3 sub-scales of SSES (see Table 1), the MVPA with these SSES scores produced similar results as that with RSES. Accordingly, for explicit self-esteem, we report only MVPA results with RSES scores.

# Results

## Behavioral Results

Self-esteem IAT scores were significantly positive (mean IAT D score = 0.69, $t(42) = 12.58$, $p < 0.001$, Cohen's $d = 1.90$). Also, the self-esteem IAT was uncorrelated with the RSES ($r = -0.07$, $p = 0.63$; a 95% confidence interval of the correlation was -0.36 to 0.24). This correlation is slightly lower, but compatible with prior findings (Hofmann, Gawronski, Gschwendner, Le, & Schmitt, 2005). The RSES was significantly correlated with each sub-scale of the SSES (see Table 1 for all correlational results). Of note, the self-esteem IAT was related neither to self-face attractiveness ratings ($r = -0.23$, $p = 0.14$) nor the appearance sub-scale of SSES ($r = -0.01$, $p = 0.94$), whereas these two measures were significantly correlated with the RSES ($r$s > 0.49, $p$s < 0.001; Table 1). Thus, any of the fMRI results reported below are unlikely to be explained by participants' perceived self-face attractiveness.

Inside the scanner, we instructed participants to press a button when luminance of an image changed. The average performance of this detection task was 96.6% for the self-image blocks and 93.8% for the scrambled-image blocks, and they were not significantly different from each other ($t(41) = 1.86$, $p = 0.07$, $d = 0.33$). Average reaction times were faster in the self-image block (431 ms) compared to the scrambled image blocks (453 ms) ($t(41) = 2.02$, $p = 0.05$, $d = 0.19$), suggesting that participants' own self-faces were more attention grabbing. Importantly, however, neither the self-esteem IAT ($r = -0.19$, $p = 0.22$) nor the RSES ($r = -0.10$, $p = 0.52$) was related to reaction times in the self-image blocks.

**fMRI Results (MVPA)**

We first defined the reward-related brain regions using Neurosynth (Yarkoni et al., 2011) (Figure 2a). These are the regions that are most preferentially related to reward (e.g., reverse inference map). Consistent with our hypothesis, activation patterns in the large reward-related ROI were robustly associated with the self-esteem IAT (correlation between predicted vs. actual scores, $r = 0.49$, $p$ value based on permutation test [$p_{\text{perm}}$] = 0.003; Figure 2b & c), thus providing unique evidence for the validity of the self-esteem IAT.[4] Furthermore, we ran the same MVPA using the data in the regions related to reward and valuation based on the prior meta-analysis (Bartra et al., 2013) and obtained a similar result ($r = 0.43$, $p_{\text{perm}} = 0.014$).

[4] To ascertain that the above result (Figure 2) is not contingent on the way we divided participants into 6 groups in the 6-fold cross-validation (i.e., 6 groups with roughly the same means and variances), we randomly allocated participants to 6 groups to run the cross-validation and repeated this step 5,000 times. The average correlation between predicted and actual self-esteem IAT scores was $r = 0.40$. Next, we ran a permutation test where we used 5,000 randomly shuffled permutations of self-esteem IAT for decoding (the scores were shuffled across all participants in every iteration). Based on the permutation test, the average correlation of $r = 0.40$ corresponds to $p_{perm} = 0.014$.

------------------- Insert Figure 2 about here -------------------

Although we selected the above two ROIs based on Neurosynth term-based meta-analysis (Figure 2a) and a meta-analysis of fMRI studies (Bartra et al., 2013), these regions are not perfectly selective to reward. Thus, it is possible that neural signals in these ROIs and implicit self-esteem were related not because these regions are involved in automatic evaluation of the self, but due to other reasons like self-processing. To examine this possibility, we ran the same MVPA with another ROI (Figure 3a) that does not include regions in the frontal cortex such as mPFC and vmPFC, both of which are implicated in self-processing (Denny et al., 2012; Northoff et al., 2006). Neural signals in the ROI predicted implicit self-esteem ($r = 0.38$, $p_{perm} = 0.026$). We also run the MVPA using only regions that are highly selective to reward (Figure 3b). Even with this conservative test (we likely discarded at least some reward-related signals by limiting our analyses to the small region), neural signals in these regions predicted implicit self-esteem ($r = 0.36$, $p_{perm} = 0.036$).

------------------- Insert Figure 3 about here -------------------

We further tested whether the self-esteem IAT could be predicted by neural signals in each of different anatomical areas, which have been implicated in reward processing. We ran the MVPA with the self-esteem IAT scores within each of the 13 reward ROIs. Self-esteem IAT scores were significantly predicted by neural signals in 3 out of the 13 ROIs (vmPFC, left pallidum, and midbrain; Table 2). Furthermore, although prediction performances did not reach the significance in the other 10 ROIs, on average, the self-esteem IAT was significantly associated with activation patterns in the 13 reward ROIs (average $r$ = 0.24, $t[12]$ = 5.42, $p_{perm}$ = 0.008; Figure 4). In contrast, neural signals in the 55 non-reward ROIs (Table 3) were, on average, unrelated to the self-esteem IAT ($t[54]$ = 2.22, $p_{perm}$ = 0.23; Figure 4). The difference between the two groups of ROIs was significant ($t[66]$ = 3.00, $p_{perm}$ = 0.046). These results indicate that self-esteem IAT scores are related to neural signals in the reward related brain regions, but not to neural signals in the non-reward related brain regions, thus further providing evidence for the validity of implicit self-esteem IAT.

------------------- Insert Figure 4 about here -------------------

**Similarity in Neural Representations between Implicit and Explicit Self-Esteem**

We repeated the same MVPA analyses using the explicit self-esteem (RSES) scores instead of the self-esteem IAT. The large reward-related ROI (Figure 2a) was not predictive of the RSES ($r$ = -0.08, $p_{perm}$ = 0.67). Prediction performances (correlations) using neural signals from the two additional reward ROIs (Figure 3) were not significant either ($rs$ < -0.03, $p_{perm}$ > 0.50). Furthermore, when we applied the MVPA in each anatomical region among 13 reward-ROIs, the average prediction performance was not significantly different from zero ($t[12]$ = 1.05, $p_{perm}$ = 0.61) and from the average performance of the 55 non-reward ROIs ($t[66]$ = 2.14, $p_{perm}$ = 0.23; Tables 2 and 3), although prediction performances were significant in 3 of 13 ROIs (i.e., vmPFC, right pallidum, left putamen; Table 2). Thus, explicit self-esteem was not associated with neural signals in the reward related areas.

Furthermore, although both the self-esteem IAT and RSES were associated with at least some of the reward ROIs at uncorrected $p_{perm} < 0.05$ level (Tables 2 and 3), among the 13 reward ROIs, the prediction performances were uncorrelated between the self-esteem IAT and RSES ($r = -0.37$, $p_{perm} = 0.24$). They were also uncorrelated across all 68 ROIs ($r = -0.06$, $p_{perm} = 0.91$). Moreover, the results showed that neural signals only in the vmPFC were commonly associated with both the self-esteem IAT and RSES (Table 2), indicating that neural signals in the vmPFC are linked with individual differences in both implicit and explicit self-esteem. However, when we computed a correlation between weight values of the self-esteem IAT and RSES, they were uncorrelated ($r = 0.11$, $p_{perm} = 0.21$), suggesting that the contribution of each voxel within the vmPFC to the prediction of the self-esteem IAT versus RSES differed.

**Exploratory MVPA Results**

Having provided the evidence supporting the validity of self-esteem IAT, we examined whether the self-esteem IAT (and also the RSES) is related to neural signals in other regions.[5] Particularly, given that self-esteem refers to how individuals view themselves, neural signals in regions involved in self-reference processing, namely mPFC and PCC (Denny et al., 2012; Northoff et al., 2006), may be related to the self-esteem IAT and/or the RSES. To test this possibility, we first ran MVPA using all voxels within the self-related ROIs (a total of 607

[5] The results reported in Table 3 address this question, at least partially. However, the table does not include all brain regions. More specifically, the following five regions do not feature in the table; (1) mPFC, (2) middle cingulate cortex (MCC), (3) posterior cingulate cortex (PCC), (4) left insula, and (5) right insula. These regions are included in the forward-inference map obtained from Neurosynth, but not in the reverse-inference map (see Methods). In other words, the five regions are consistently activated by reward, but activation in each region is not selective to reward (thus not informative to our main research question). For the sake of completeness, we ran MVPA using neural signals in each region. Neural signals in the mPFC (Frontal_Sup_Medial_R and Frontal_Sup_Medial_L masks from the WFU pickatlas toolbox; a total of 1,548 voxels) and left insula (507 voxels) significantly predicted the self-esteem IAT (mPFC, $r = 0.46$, $p_{perm} = 0.008$; left insula, $r = 0.39$, $p_{perm} = 0.022$ [uncorrected for multiple comparisons]). The remaining three regions did not predict the self-esteem IAT ($0.00 < rs < 0.23$, $p_{perm} > 0.15$). None of the five regions significantly predicted the RSES ($-0.22 < rs < 0.08$, $p_{perm} > 0.35$).

voxels; Figure 5a). Interestingly, we found that neural signals in the self-related brain regions significantly predicted both the self-esteem IAT ($r$ = 0.50, $p_{perm}$ = 0.005; Figure 5b) and the RSES ($r$ = 0.39, $p_{perm}$ = 0.023; Figure 5c). We also examined whether neural signals in each of the mPFC and PCC ROIs predicted implicit and explicit self-esteem. Indeed, the self-esteem IAT was significantly predicted by neural signals in the mPFC ($r$ = 0.49, $p_{perm}$ = 0.009), and the PCC showed a similar trend ($r$ = 0.31, $p_{perm}$ = 0.065). In contrast, explicit self-esteem was not predicted by neural signals in either mPFC ($r$ = 0.18, $p_{perm}$ = 0.18) or PCC ($r$ = -0.12, $p_{perm}$ = 0.67). Furthermore, although neural signals in the self-related ROI (607 voxels; Figure 5a) predicted both the self-esteem IAT and RSES, weight values of the self-esteem IAT and RSES were uncorrelated with each other, indicating that they are represented differently in these regions ($r$ = -0.06, $p_{perm}$ = 0.67).[6]

------------------ Insert Figure 5 about here -------------------

Another possibility is that implicit (and explicit) self-esteem may modulate how individuals view their faces, and thus may be related to neural signals in regions involved in face processing such as fusiform gyrus. Consistent with this possibility, an fMRI study has demonstrated that fusiform activation for White faces relative to Black faces was significantly

[6] We further tested whether we could better predict implicit self-esteem by aggregating neural signals from both the reward- and self-related ROIs (Figures 2a & 5a). We combined the two ROIs (a total of 3,189 voxels) and ran MVPA. The result showed that the correlation between actual and predicted self-esteem IAT scores was $r$ = 0.50 ($p_{perm}$ = 0.005), which is compatible to what we found using the large reward ROI only ($r$ = 0.49; Figure 2). Thus, combining the two ROIs (reward and self ROIs) did not increase the prediction performance. However, it should be noted that the size of correlation we found in our main analysis ($r$ = 0.49) seems to be already at its ceiling; that is, based on the power analysis we reported above, we estimated the effect size to be $r$ = 0.392. Hence, it is theoretically difficult to demonstrate the additive nature of signals from the two ROIs in predicting implicit self-esteem.

correlated with implicit prejudice (i.e., race IAT scores; Cunningham et al., 2004). Further, more recent MVPA studies indicate that neural signals in fusiform face area (FFA) in response to faces are modulated depending on implicit attitudes (Brosch, Bar-David, & Phelps, 2013) or stereotypes (Stolier & Freeman, 2016). However, our results showed that activations in both left and right fusiform gyrus were unassociated with the self-esteem IAT (left $r$ = 0.21, $p_{perm}$ = 0.17; right $r$ = 0.26, $p_{perm}$ = 0.12; Table 3), although both correlations were in a positive direction. The RSES was also unassociated with activations in fusiform gyrus (left $r$ = -0.46, $p_{perm}$ = 0.99; right $r$ = -0.09, $p_{perm}$ = 0.65).[7]

**fMRI Results (Univariate Analysis)**

We further tested whether the self-esteem IAT and RSES were linearly related to the level of univariate activity in reward-related brain regions. In the reward ROI (Figure 2a), no region was significantly related, either positively or negatively, to either the self-esteem IAT or RSES. Similarly, there was no significant region outside of the ROI for either the self-esteem IAT or RSES. The results suggest that the level of univariate activity in response to self-face is unrelated to implicit and explicit self-esteem.

## Discussion

We aimed to provide unique evidence for the validity of an implicit self-esteem measure using neuroimaging combined with a machine learning technique, MVPA. Our findings indicate that implicit self-esteem, as measured by the IAT, is associated with neural activation patterns automatically evoked by passive viewing of self-face in the reward-related regions (Figures 2a, 3a, and 3b) as well as in 13 reward-related anatomical ROIs (Table 2 and Figure 4), but not in

[7] We also defined face selective regions in ventral occipito-temporal cortex in two ways: using (1) the self-face versus scrambled-image contrast, and (2) Neurosynth term-based meta-analysis with the term "face." We ran MVPA employing neural signals in each of the two ROIs, but did not obtain significant result for either the self-esteem IAT or RSES.

non-reward related areas (Table 3 and Figure 4). Thus, although in prior research (Falk & Heine, 2015) implicit self-esteem measures were found to be unrelated to personality or attitude measures (i.e., had low convergent and predictive validity), in our study self-esteem IAT scores were robustly associated with (i.e., predictive of) neural signals in a way that is consistent with the conceptualization of implicit self-esteem as an automatic attitude toward the self (Greenwald & Banaji, 1995; Sedikides & Gregg, 2003). The literature has indicated that attractive faces activate reward-related brain areas (Cloutier, Heatherton, Whalen, & Kelley, 2008; O'Doherty et al., 2003), and that face attractiveness can be decoded from neural signals in vmPFC (Pegors, Kable, Chatterjee, & Epstein, 2015). However, given that IAT scores were unrelated to both perceived self-face attractiveness and the SSES sub-scale of appearance (while both being significantly related to explicit self-esteem scores, i.e., RSES), our results are unlikely to be mediated by individual difference in perceived self-face attractiveness.

Our study provides important and independent evidence supporting the validity of the self-esteem IAT, and offers a unique insight into the debate on the validity of implicit self-esteem measures. For example, although prior results suggest that implicit self-esteem measures lack convergent validity (Bosson et al., 2000; Falk et al., 2015; Rudolph et al., 2008), the present findings demonstrate that the low convergent validity is likely due to low validity of other implicit measures, but not the IAT. One task of future research would be to examine the validity of other implicit self-esteem measures (e.g., name-letter task; Nuttin, 1985) using the neuroimaging approach.

Similarly, as stated earlier, the low predictive validity of implicit self-esteem measures may be due to biases in selecting criterion variables, which is likely due to lack of clear understanding of what implicit self-esteem is. Nonetheless, some research (Cvencek, Greenwald, & Meltzoff, 2016; Greenwald et al., 2002) has shown that implicit self-esteem, gender identity, and gender attitude (all measured by IAT) are related to each other in a manner consistent with balanced identity theory (Greenwald et al., 2002), illustrating that the self-esteem IAT can predict other implicit attitudes that are selected on the basis of firm theoretical background. Interestingly, our fMRI results indicated that neural signals in the regions involved in self-processing (Figure 5a) were associated with both implicit and explicit self-esteem, thus suggesting that both implicit and explicit self-esteem may be related to the proclivity for automatic engagement in self-reference (Gregg, Mahadevan, & Sedikides, 2017; Rogers, Kuiper, & Kirker, 1977). Yet, we noted that, just like any other brain regions, the mPFC and PCC are not perfectly selective to self-processing, and our findings may be accounted for, at least partially, by other processes. For example, as discussed above, the mPFC is implicated in reward-processing (Kable & Glimcher, 2007; Knutson, Fong, Bennett, Adams, & Hommer, 2003). Similarly, the PCC is implicated in episodic memory (Hassabis, Kumaran, & Maguire, 2007). Thus, future behavioral studies should test this unique hypothesis (i.e., the link between implicit self-esteem and self-reference processing) in order to provide further insight into what the self-esteem IAT is measuring.

We found not only that implicit and explicit self-esteem were linked to neural signals in self-related regions (Figure 5), but also that they were linked so in different ways. Implicit self-esteem was represented in each of the two self-related ROIs independently (although evidence for the PCC was weak [i.e., $p_{perm} = 0.065$]), whereas explicit self-esteem was *collectively* represented in the mPFC and PCC ROIs (i.e., alone the ROIs could not predict explicit self-esteem). The result may suggest that two distinct processes interact with each other and determine explicit evaluation of the self (explicit self-esteem). A fitting analogy may be the Associative–Propositional Evaluation (APE) model of attitudes (Gawronski & Bodenhausen, 2006), which postulates that implicit and explicit evaluations are the outcomes of two distinct processes: (1) associative, and (2) propositional. The APE model states that, although implicit evaluations depend on associative processes (i.e., automatically activated associations), explicit evaluations depend on activated associations (associative processes) and their validation according to cognitive consistency principles (propositional processes). It is, of course, rather simplistic to regard the associative and propositional processes of the APE model as mapping directly onto the mPFC and PCC, respectively. Yet, it is possible that explicit self-esteem is determined by a similar interaction process between two (unspecified) distinct processes.

Our study also evidence, albeit indirect, for the divergent validity of implicit and explicit self-esteem. Explicit self-esteem was not associated with neural signals in the large-reward related ROI (Figure 2a). Furthermore, neural representations of implicit and explicit self-esteem are largely distinct on a local level (i.e., within the vmPFC ROI, within the self-related ROI [Figure 5a], and across 13 reward-ROIs) as well as on a global level (i.e., across all 68 ROIs; Tables 2 and 3), supporting the idea that implicit and explicit self-esteem are distinct constructs (Greenwald & Farnham, 2000; Jordan, Logel, Spencer, Zanna, & Whitfield, 2009). This finding, though, should be interpreted with caution. The less clear relation between neural signals in the reward related areas and explicit self-esteem is probably due to the use of automatic brain activations in response to self-face for prediction (i.e., passive-viewing), a practice less likely to be linked with conscious and reflective self-evaluation (explicit self-esteem). Given previous studies demonstrating a link between explicit self-esteem and neural activities in reward-related brain regions (Chavez & Heatherton, 2015; Frewen et al., 2013; Oikawa et al., 2012), it is plausible that these regions play a key role in explicit self-esteem as well as implicit self-esteem. Thus, future research would do well to test whether neural signals in the reward-related regions, while participants are engaging in explicit evaluations of self (e.g., self-reference task), can predict individual differences in explicit self-esteem and differences/similarities in how implicit and explicit self-esteem are represented in these regions.

We based the study's design on findings that activity in the reward related brain regions as a response to an object reflects participants' preference for that object (Izuma et al., 2017; Lebreton et al., 2009; Levy et al., 2011; Smith et al., 2014; Tusche et al., 2010). One might argue, however, that evidence could have been stronger, if we demonstrated that a decoder of preference for non-social reward objects (e.g., food) could predict implicit self-esteem (i.e., a more direct link between activity in the reward related areas and neural signals as a response to self-face). Here, we would first train the prediction model on responses to a food, then apply this model to neural responses to one's own face, and finally test if it can predict self-esteem IAT scores. Although such a demonstration would have been ideal, this proposal would rely on the assumption that preferences for non-social objects and attitudes toward the self are represented in a similar manner in the brain. Such an assumption is empirically unsupported. Previous neurophysiological studies with monkeys and rats established that largely distinct populations of striatal neurons encode reward values of different types of reward (e.g., juice vs. drug rewards; Bowman, Aigner, & Richmond, 1996; Carelli, Ijames, & Crumling, 2000; Carelli & Wondolowski, 2003; Robinson & Carelli, 2008). A recent MVPA study also indicated that, although there may exist a population of neurons that encode both social and non-social rewards, these two types of rewards are processed in largely distinct neural circuits (Wake & Izuma, 2017).

We recruited only young female individuals in Western culture. It is interesting and important to test whether the findings can be replicated in males or individuals from different cultures. In addition to testing the validity of the self-esteem IAT, our study also afforded a novel insight into what implicit self-esteem (as measured by the IAT) is by demonstrating an association between neural signals in self-processing regions (i.e., mPFC and PCC) and implicit (and explicit) self-esteem. Prior research (Kitayama & Uchida, 2003; Yamaguchi et al., 2007) showed that, whereas people in Western countries tend to have higher explicit self-esteem than those in East-Asian countries, both cultures manifest the same level of implicit self-esteem (for a review, see: Sedikides, Gaertner, & Cai, 2015). Future empirical efforts could be directed toward addressing similarities/differences between Western and Eastern cultures in terms of neural representations of implicit and explicit self-esteem.

In conclusion, our study highlights the utility of neuroimaging methods combined with the MVPA to test a psychological hypothesis. MVPA is more suitable for identifying complex neural representations of higher cognitive processes such as self-esteem than conventional fMRI data analysis. Although the present study focused on testing the validity of the self-esteem IAT, the same approach can be applied to any explicit or implicit measure, as long as there is a sensible hypothesis about brain regions involved in a measured psychological construct (e.g., self-esteem [attitude toward the self] = reward-related brain regions). Thus, a machine learning (MVPA) approach could provide not only unique insight into the validity of psychological measures, but also advance psychological theories in a way that goes above and beyond existing behavioral measures.

## Acknowledgements

We thank Ellen Kingsley and Olivia Rogerson for their assistance in data collection and the staff of the Neuroimaging Centre for their assistance in conducting the fMRI. K.S. was supported by JSPS Kakehnhi Grant Number 17H04789.

**References**

Ahn, W. Y., Kishida, K. T., Gu, X., Lohrenz, T., Harvey, A., Alford, J. R., . . . Montague, P. R. (2014). Nonpolitical images evoke neural predictors of political ideology. *Current Biology, 24*(22), 2693–2699.

Alpaydin, E. (Ed.) (2014). *Introduction to machine learning (3rd ed)*. Cambridge, MA: The MIT Press.

Amodio, D. M. (2010). Can neuroscience advance social psychological theory? Social neuroscience for the behavioral social psychologist. *Social Cognition, 28*(6), 695-716.

Ariely, D., & Berns, G. S. (2010). Neuromarketing: the hope and hype of neuroimaging in business. *Nature Reviews Neuroscience, 11*(4), 284-292. doi:10.1038/nrn2795

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavioral Research Methods, 46*(3), 668-688. doi:10.3758/s13428-013-0410-6

Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage, 76*, 412-427. doi:10.1016/j.neuroimage.2013.02.063

Bjork, J. M., Knutson, B., Fong, G. W., Caggiano, D. M., Bennett, S. M., & Hommer, D. W. (2004). Incentive-elicited brain activation in adolescents: similarities and differences from young adults. *Journal of Neuroscience, 24*(8), 1793-1802. doi:10.1523/JNEUROSCI.4862-03.2004

Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions, questionable assumptions and post hoc rationalizations: Will the real IAT, please stand up? *Journal of Experimental Social Psychology, 43*(3), 399-409.

doi:10.1016/j.jesp.2006.10.019

Blanton, H., Klick, J., Mitchell, G., Jaccard, J., Mellers, B., & Tetlock, P. E. (2009). Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT. *Journal of Applied Psychology, 94*(3), 567-582. doi:10.1037/a0014665

Bleidorn, W., Arslan, R. C., Denissen, J. J., Rentfrow, P. J., Gebauer, J. E., Potter, J., & Gosling, S. D. (2016). Age and gender differences in self-esteem-A cross-cultural window. *Journal of Personality and Social Psychology, 111*(3), 396-410. doi:10.1037/pspp0000078

Bosson, J. K., Swann, 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. doi:10.1037//0022-3514.79.4.631

Bowman, E. M., Aigner, T. G., & Richmond, B. J. (1996). Neural signals in the monkey ventral striatum related to motivation for juice and cocaine rewards. *Journal of Neurophysioly, 75*(3), 1061-1073.

Brosch, T., Bar-David, E., & Phelps, E. A. (2013). Implicit race bias decreases the similarity of neural representations of black and white faces. *Psychological Science, 24*(2), 160-166. doi:10.1177/0956797612451465

Buhrmester, M. D., Blanton, H., & Swann, W. B. (2011). Implicit self-mesteem: Nature, measurement, and a new way forward. *Journal of Personality and Social Psychology, 100*(2), 365-385. doi:10.1037/A0021341

Carelli, R. M., Ijames, S. G., & Crumling, A. J. (2000). Evidence that separate neural circuits in the nucleus accumbens encode cocaine versus "natural" (water and food) reward. *Journal of Neuroscience, 20*(11), 4255-4266.

Carelli, R. M., & Wondolowski, J. (2003). Selective encoding of cocaine versus natural rewards by nucleus accumbens neurons is not related to chronic drug exposure. *Journal of Neuroscience, 23*(35), 11214-11223.

Chavez, R. S., & Heatherton, T. F. (2015). Multimodal frontostriatal connectivity underlies individual differences in self-esteem. *Social Cognitive and Affective Neuroscience, 10*(3), 364-370. doi:10.1093/scan/nsu063

Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2017). Neural population decoding reveals the intrinsic positivity of the self. *Cerebral Cortex, 27*(11), 5222-5229. doi:10.1093/cercor/bhw302

Cloutier, J., Heatherton, T. F., Whalen, P. J., & Kelley, W. M. (2008). Are attractive people rewarding? Sex differences in the neural substrates of facial attractiveness. *Journal of Cognitive Neuroscience, 20*(6), 941-951. doi:10.1162/jocn.2008.20062

Cohen, J. R., Asarnow, R. F., Sabb, F. W., Bilder, R. M., Bookheimer, S. Y., Knowlton, B. J., & Poldrack, R. A. (2010). Decoding developmental differences and individual variability in response inhibition through predictive analyses across individuals. *Frontiers in Human Neuroscience, 4*, 47. doi:10.3389/fnhum.2010.00047

Cunningham, W. A., Johnson, M. K., Raye, C. L., Chris Gatenby, J., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science, 15*(12), 806-813. doi:10.1111/j.0956-7976.2004.00760.x

Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2016). Implicit measures for preschool children confirm self-esteem's role in maintaining a balanced identity. *Journal of Experimental Social Psychology, 62*, 50-57. doi:10.1016/j.jesp.2015.09.015

Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience, 24*(8), 1742-1752. doi:10.1162/jocn_a_00233

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems 9*(9), 155-161.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170-177. doi:10.1037/1082-989x.1.2.170

Falk, C. F., & Heine, S. J. (2015). What is implicit self-esteem, and does it vary across cultures? *Personality and Social Psychology Review, 19*(2), 177-198. doi:10.1177/1088868314544693

Falk, C. F., Heine, S. J., Takemura, K., Zhang, C. X., & Hsu, C. W. (2015). Are implicit self-esteem measures valid for assessing individual and cultural differences? *Journal of Personality, 83*(1), 56-68. doi:10.1111/jopy.12082

Frewen, P. A., Lundberg, E., Brimson-Theberge, M., & Theberge, J. (2013). Neuroimaging self-esteem: a fMRI study of individual differences in women. *Social Cognitive and Affective Neuroscience, 8*(5), 546-555. doi:10.1093/scan/nss032

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2*, 189-210. doi:10.1002/hbm.460020402

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692-731. doi:10.1037/0033-2909.132.5.692

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27. doi:10.1037//0033-295x.102.1.4

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review, 109*(1), 3-25. doi:10.1037//0033-295x.109.1.3

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. doi:10.1037//0022-3514.79.6.1022

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197-216. doi:10.1037/0022-3514.85.2.197

Gregg, A. P., Mahadevan, N., & Sedikides, C. (2017). The SPOT effect: People spontaneously prefer their own theories. *The Quarterly Journal of Experimental Psychology, 70*(6), 996-1010. doi:10.1080/17470218.2015.1099162

Gregg, A. P., & Sedikides, C. (2010). Narcissistic fragility: Rethinking its links to explicit and implicit self-esteem. *Self and Identity, 9*(2), 142-161. doi:10.1080/15298860902815451

Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology, 35*(1), 4-26. doi:10.1038/npp.2009.129

Hariri, A. R., Brown, S. M., Williamson, D. E., Flory, J. D., de Wit, H., & Manuck, S. B. (2006). Preference for immediate over delayed rewards is associated with magnitude of ventral striatal activity. *Journal of Neuroscience, 26*(51), 13213-13217.

doi:10.1523/JNEUROSCI.3446-06.2006

Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *Journal of Neuroscience, 27*(52), 14365-14374. doi:10.1523/JNEUROSCI.4549-07.2007

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience, 7*(7), 523-534. doi:10.1038/nrn1931

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), 895-910. doi:10.1037//0022-3514.60.6.895

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*(10), 1369-1385. doi:10.1177/0146167205275613

Honekopp, J. (2006). Once more: is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perceptiion and Performance, 32*(2), 199-209. doi:10.1037/0096-1523.32.2.199

Izuma, K. (2015). Social reward. In A. W. Toga (Ed.), *Brain mapping: An encyclopedic reference* (Vol. 3, pp. 21-23). Oxford, UK: Elsevier.

Izuma, K., Shibata, K., Matsumoto, K., & Adolphs, R. (2017). Neural predictors of evaluative attitudes towards celebrities. *Social Cognitive and Affective Neuroscience, 12*, 382-390. doi:10.1093/scan/nsw135

Jimura, K., & Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia, 50*(4), 544-552. doi:10.1016/j.neuropsychologia.2011.11.007

Jordan, C. H., Logel, C., Spencer, S. J., Zanna, M. P., & Whitfield, M. L. (2009). The heterogeneity of self-esteem. In R. E. Petty, R. H. Fazio, & P. Brinol (Eds.), *Attitudes: Insight from the new implicit measures*. New York, NY: Psychology Press.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience, 10*(12), 1625-1633. doi:10.1038/nn2007

Kaplan, J. T., Aziz-Zadeh, L., Uddin, L. Q., & Iacoboni, M. (2008). The self across the senses: an fMRI study of self-face and self-voice recognition. *Social Cognitive Affective Neuroscience, 3*(3), 218-223. doi:10.1093/scan/nsn014

Kitayama, S., & Uchida, Y. (2003). Explicit self-criticism and implicit self-regard: Evaluating self and friend in two cultures. *Journal of Experimental Social Psychology, 39*(5), 476-482. doi:10.1016/S0022-1031(03)00026-X

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125*(4), 470-500. doi:10.1037//0033-2909.125.4.470

Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience, 21*(16), RC159.

Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. M., & Hommer, D. (2003). A region of mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with rapid event-related fMRI. *Neuroimage, 18*(2), 263-272.

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *Journal of Neuroscience, 25*(19), 4806-4812. doi:10.1523/JNEUROSCI.0642-05.2005

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection.* Paper presented at the International Joint Conference on Artificial Intelligence, San Fransisco, CA.

Kolling, N., Behrens, T., Wittmann, M. K., & Rushworth, M. (2016). Multiple signals in anterior cingulate cortex. *Current Opinion in Neurobiology, 37*, 36-43. doi:10.1016/j.conb.2015.12.007

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., & Ono, T. (2001). Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature, 412*(6846), 546-549. doi:10.1038/35087595

Krause, S., Back, M. D., Egloff, B., & Schmukle, S. C. (2011). Reliability of Implicit Self-esteem Measures Revisited. *European Journal of Personality, 25*(3), 239-251. doi:10.1002/Per.792

Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron, 64*(3), 431-439. doi:10.1016/j.neuron.2009.09.040

Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011). Choice from non-choice: predicting consumer preferences from blood oxygenation level-dependent signals obtained during passive viewing. *Journal of Neuroscience, 31*(1), 118-125. doi:10.1523/JNEUROSCI.3214-10.2011

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets.

*Neuroimage, 19*(3), 1233-1239.

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience, 8*(3), 285-299. doi:10.1093/scan/nsr090

Mizuhiki, T., Richmond, B. J., & Shidara, M. (2012). Encoding of reward expectation by monkey anterior insular neurons. *Journal of Neurophysioly, 107*(11), 2996-3007. doi:10.1152/jn.00282.2011

Nishijo, H., Ono, T., & Nishino, H. (1988). Single neuron responses in amygdala of alert monkey during complex sensory stimulation with affective significance. *Journal of Neuroscience, 8*(10), 3570-3583.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*(9), 424-430. doi:10.1016/j.tics.2006.07.005

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain: a meta-analysis of imaging studies on the self. *Neuroimage, 31*(1), 440-457.

Nuttin, J. M., Jr. (1985). Narcissism beyond Gestalt and awareness: the name letter effect. *European Journal of Social Psychology, 15*, 353-361.

O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty in a smile: the role of medial orbitofrontal cortex in facial attractiveness. *Neuropsychologia, 41*(2), 147-155.

Oikawa, H., Sugiura, M., Sekiguchi, A., Tsukiura, T., Miyauchi, C. M., Hashimoto, T., . . . Kawashima, R. (2012). Self-face evaluation and self-esteem in young females: an fMRI study using contrast effect. *Neuroimage, 59*(4), 3668-3676. doi:10.1016/j.neuroimage.2011.10.098

Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage, 49*(3), 1943-1948. doi:10.1016/J.Neuroimage.2009.02.047

Pegors, T. K., Kable, J. W., Chatterjee, A., & Epstein, R. A. (2015). Common and unique representations in pFC for face and place attractiveness. *Journal of Cognitive Neuroscience, 27*(5), 959-973. doi:10.1162/jocn_a_00777

Peirce, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8-13. doi:10.1016/j.jneumeth.2006.11.017

Pliner, P., Chaiken, S., & Flett, G. L. (1990). Gender differences in concern with body weight and physical appearance over the life-span. *Personality and Social Psychology Bulletin, 16*(2), 263-273. doi:10.1177/0146167290162007

Robinson, D. L., & Carelli, R. M. (2008). Distinct subsets of nucleus accumbens neurons encode operant responding for ethanol versus water. *European Journal of Neuroscience, 28*(9), 1887-1894. doi:10.1111/j.1460-9568.2008.06464.x

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology, 35*(9), 677-688. doi:10.1037/0022-3514.35.9.677

Rosenberg, M. (1965). *Society and the adolescent self-image.* Princeton, NJ: Princeton University Press.

Rudolph, A., Schroder-Abe, M., Schutz, A., Gregg, A. P., & Sedikides, C. (2008). Through a glass, less darkly? Reassessing convergent and discriminant validity in measures of implicit self-esteem. *European Journal of Psychological Assessment, 24*(4), 273-281. doi:10.1027/1015-5759.24.4.273

Sapountzis, P., Schluppeck, D., Bowtell, R., & Peirce, J. W. (2010). A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. *Neuroimage, 49*(2), 1632-1640. doi:10.1016/j.neuroimage.2009.09.066

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Review, 95*(3), 853-951. doi:10.1152/physrev.00023.2014

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience, 13*(3), 900-913.

Sedikides, C., Gaertner, L., & Cai, H. (2015). On the panculturality of self-enhancement and self-protection motivation: The case for the universality of self-esteem. *Advances in Motivation Science, 2*, 185-241. doi:10.1016/bs.adms.2015.04.002

Sedikides, C., & Gregg, A. P. (2003). Portaits of the self. In M. A. Hogg & J. Cooper (Eds.), *Sage handbook of social psychology* (pp. 110-138). London, UK: Sage Publications.

Sescousse, G., Caldu, X., Segura, B., & Dreher, J. C. (2013). Processing of primary and secondary rewards: a quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews, 37*(4), 681-696. doi:10.1016/j.neubiorev.2013.02.002

Shibata, K., Watanabe, T., Kawato, M., & Sasaki, Y. (2016). Differential activation patterns in the same brain region led to opposite emotional states. *PLoS Biology, 14*(9), e1002546. doi:10.1371/journal.pbio.1002546

Smith, A., Bernheim, B. D., Camerer, C. F., & Rangel, A. (2014). Neural activity reveals preferences without choices. *American Economic Journal-Microeconomics, 6*(2), 1-36. doi:10.1257/Mic.6.2.1

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience, 19*(6), 795-797. doi:10.1038/nn.4296

Sugiura, M., Kawashima, R., Nakamura, K., Okada, K., Kato, T., Nakamura, A., . . . Fukuda, H. (2000). Passive and active recognition of one's own face. *Neuroimage, 11*(1), 36-48. doi:10.1006/nimg.1999.0519

Toga, A. W., & Thompson, P. M. (2003). Mapping brain asymmetry. *Nature Reviews Neuroscience, 4*(1), 37-48. doi:10.1038/nrn1009

Tusche, A., Bode, S., & Haynes, J. D. (2010). Neural responses to unattended products predict later consumer choices. *Journal of Neuroscience, 30*(23), 8024-8031. doi:10.1523/JNEUROSCI.0064-10.2010

Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience, 12*(10), 1558-1564. doi:10.1093/scan/nsx092.

Will, G., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *eLife, 6*, e28098. doi:10.7554/eLife.28098

Wu, C. C., Bossaerts, P., & Knutson, B. (2011). The affective impact of financial skewness on neural activity and choice. *PLoS One, 6*(2), e16838. doi:10.1371/journal.pone.0016838

Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K., . . . Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological*

*Science, 18*(6), 498-500. doi:10.1111/j.1467-9280.2007.01928.x

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods, 8*(8), 665-U695. doi:10.1038/Nmeth.1635

**Figures**

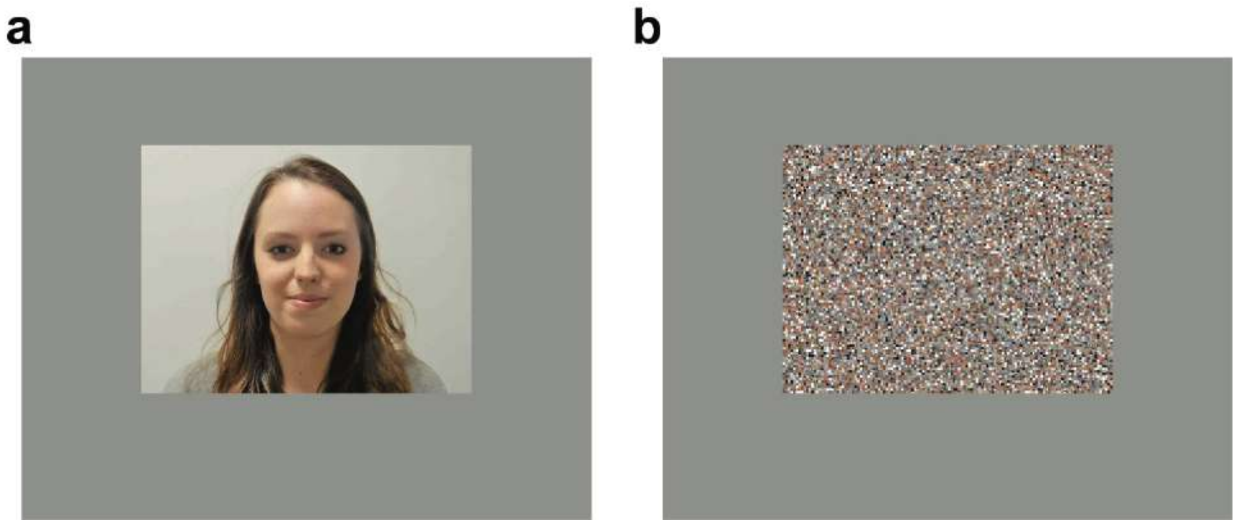


**Figure 1.** Examples of stimuli presented during fMRI scanning. Inside an fMRI scanner, a participant viewed 4 images of the self (**a**) or 4 scrambled images (**b**) in each block.

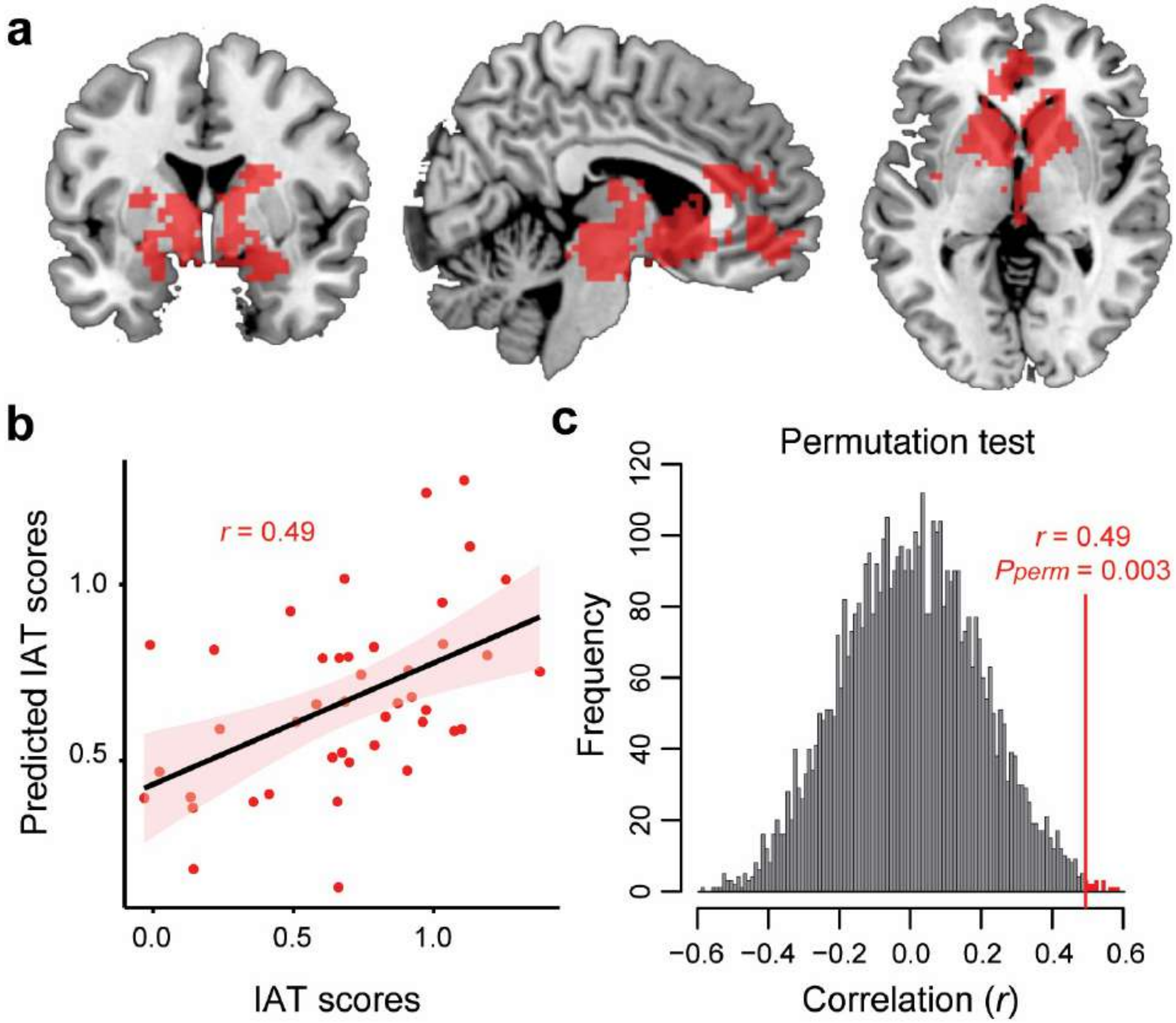


**Figure 2.** (**a**). A large reward-related ROI defined using Neurosynth (a total of 2,696 voxels). Left: coronal view (y = 0). Middle: sagittal view (x = 6). Right: Axial view (z = 0). (**b**). A correlation between participants' self-esteem IAT scores and predicted scores based on neural signals in the ROI. (**c**). A histogram showing the distributions of correlation coefficients between actual and predicted IAT scores with randomly permutated data (5,000 times). The correlation with actual data was significant at $p_{perm} = 0.003$.

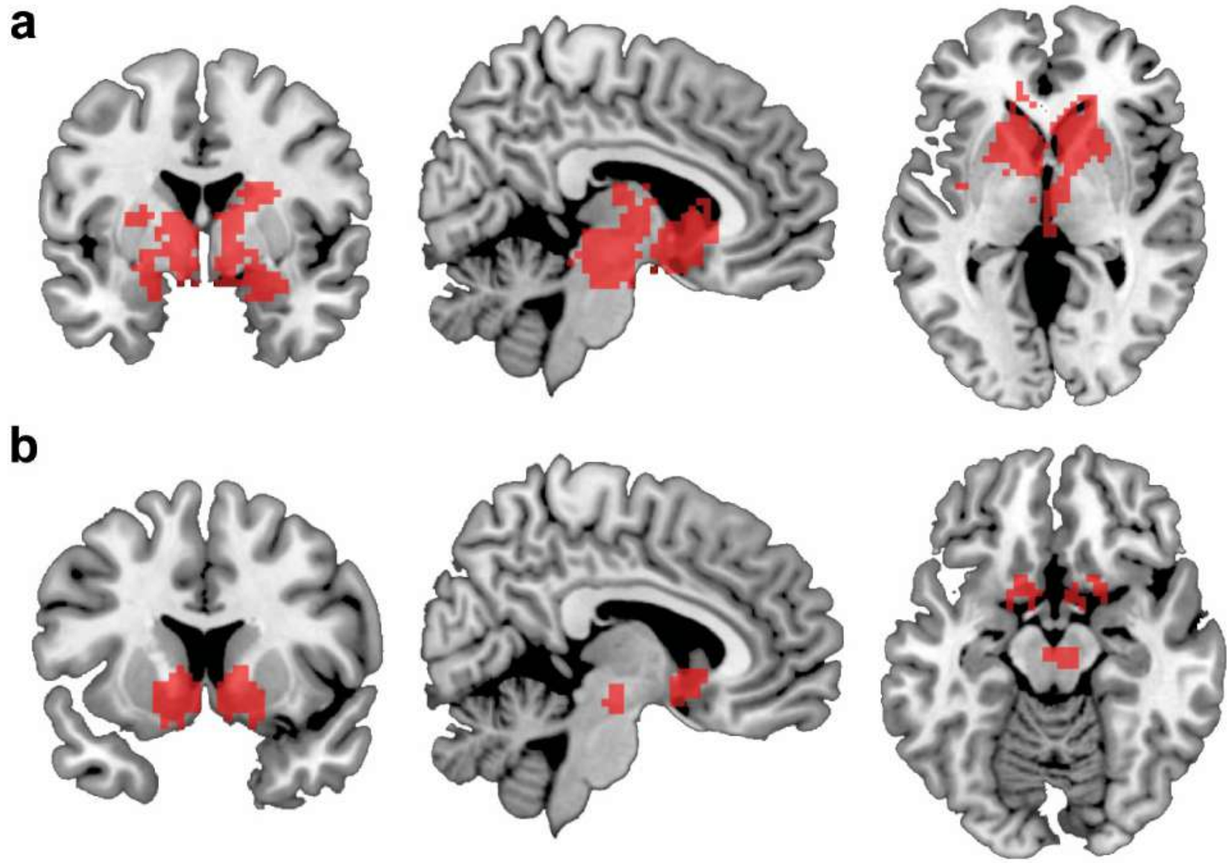


**Figure 3.** Two Additional Reward ROIs. (a) Anatomical structures in the frontal cortex (i.e., mPFC, vmPFC, ACC) were removed from the large reward ROI (Figure 2a). There are a total of 2,179 voxels. Left: coronal view (y = 0). Middle: sagittal view (x = 6). Right: Axial view (z = 0). (b) Regions highly selective to reward obtained from Neurosynth (a term-based meta-analysis with the term "Reward" and thresholded at z-score = 10). The ROI consists of bilateral ventral striatum (nucleus accumbens) and midbrain (a total of 343 voxels). Left: coronal view (y = 10). Middle: sagittal view (x = 6). Right: Axial view (z = -14).

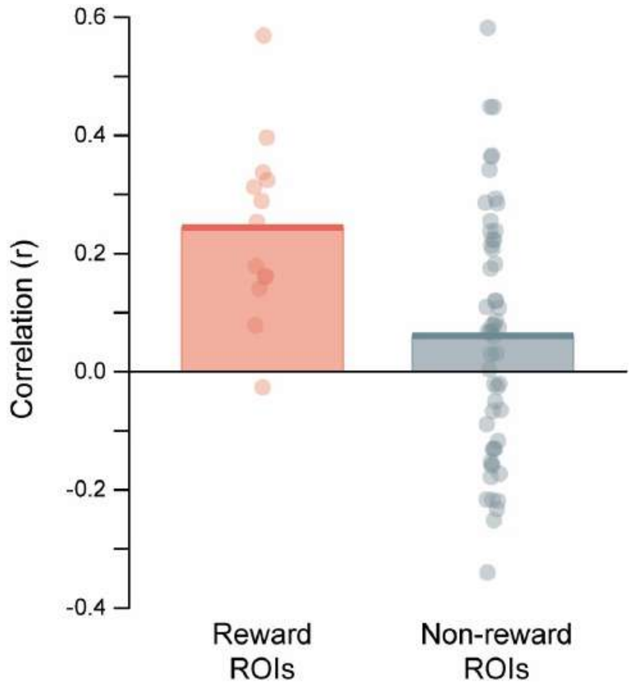


**Figure 4.** Average prediction performance (correlation between actual and predicted implicit self-esteem) in each of two groups of ROIs; 1) the 13 reward ROIs (left), and 2) 55 non-reward ROIs (right). See also Table 1. Note that the figure is based on original correlation values, although we conducted statistical tests on Fisher-z transformed values.

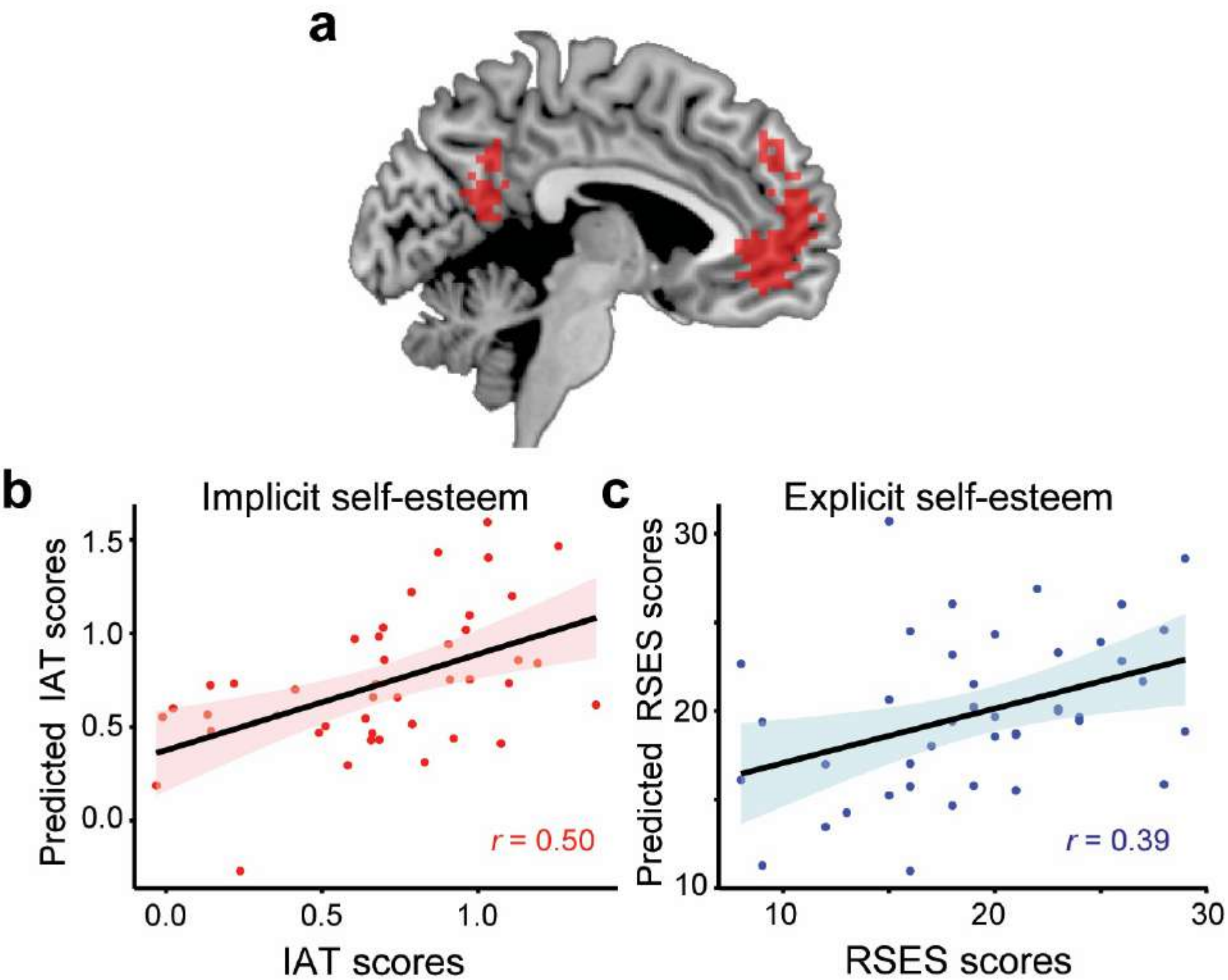


**Figure 5.** (a) Self-related ROI defined by Neurosynth (x = -5). The self-ROI consists of mPFC and PCC (a total of 607 voxels). (b) A correlation between participants' self-esteem IAT scores and predicted IAT scores based on neural signals in the ROI. (c) A correlation between participants' RSES scores and predicted RSES scores based on neural signals in the ROI.