

This is a repository copy of *Neural activity in the reward-related brain regions predicts implicit self-esteem: A novel validity test of psychological measures using neuroimaging.*

White Rose Research Online URL for this paper:
<https://eprints.whiterose.ac.uk/126293/>

Version: Accepted Version

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>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

1 © 2018, American Psychological Association. This paper is not the copy of record and may
2 not exactly replicate the final, authoritative version of the article. Please do not copy or cite
3 without authors permission. The final article will be available, upon publication, via its
4 DOI: 10.1037/pspa0000114

5
6
7 **Neural Activity in the Reward-Related Brain Regions Predicts Implicit Self-Esteem:**
8 **A Novel Validity Test of Psychological Measures Using Neuroimaging**

9
10 Keise Izuma, Kate Kennedy, and Alexander Fitzjohn

11 University of York, UK

12
13 Constantine Sedikides

14 University of Southampton, UK

15
16 Kazuhisa Shibata

17 Nagoya University, Japan

18
19
20 Corresponding author: Keise Izuma, Department of Psychology, University of York, Heslington,
21 York, YO10 5DD, UK; Tel: +44 (0)1904 323167; Email: keise.izuma@york.ac.uk

22

23

24 Abstract

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

44

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

46 Neural Activity in the Reward-Related Brain Regions Predicts Implicit Self-Esteem:**47 A Novel Validity Test of Psychological Measures Using Neuroimaging**

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

60 It is difficult, however, to make a definitive contribution to that debate, because validity
61 has been assessed in reference to other imperfect behavioral measures. For example, Falk et al.
62 (2015) collected nine implicit measures of self-esteem from three groups of participants (Euro-
63 Canadians, Asian-Canadians, Japanese). The implicit measures were uncorrelated with each
64 other across all three groups, demonstrating the low convergent validity of implicit self-esteem
65 measures. However, we cannot conclude from these results that all implicit self-esteem measures
66 are invalid: even if one measure was perfectly reliable and valid, no correlation would emerge in
67 the case in which all other measures were invalid.

68 Similarly, the low predictive validity of implicit self-esteem measures found in prior

69 research may be due to biases in selecting criterion variables. Researchers have typically selected
70 criterion variables based on understanding of what explicit self-esteem is (Bosson et al., 2000;
71 Falk et al., 2015). As a consequence, almost all criterion variables have been strongly correlated
72 with explicit self-esteem measures, but not with implicit self-esteem measures (Bosson et al.,
73 2000; Falk et al., 2015; for a review, see Buhrmester et al., 2011). Given the divergent validity of
74 implicit and explicit self-esteem (Bosson et al., 2000; Buhrmester et al., 2011; Falk et al., 2015;
75 Greenwald & Farnham, 2000; Rudolph et al., 2008), this literature may not be a fair test of the
76 predictive validity of implicit self-esteem measures. Stated otherwise, lack of predictive validity
77 may simply reflect unclarities in the definition of implicit self-esteem.

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

88 More specifically, we test whether self-esteem IAT scores are robustly associated with
89 neural activation in the reward-related brain regions (Bartra, McGuire, & Kable, 2013; Kolling,
90 Behrens, Wittmann, & Rushworth, 2016; Schultz, 2015; Sescousse, Caldu, Segura, & Dreher,
91 2013) in response to self-face—arguably, the most obvious, immediate, and authentic

92 representation of the self. Previous neuroimaging studies demonstrated that incidental
93 preferences or attitudes toward various stimuli are automatically represented (i.e., without asking
94 participants to report how they feel about the stimuli) in the reward-related areas, such as
95 striatum and ventromedial prefrontal cortex (vmPFC) (Izuma, Shibata, Matsumoto, & Adolphs,
96 2017; Lebreton, Jorge, Michel, Thirion, & Pessiglione, 2009; Levy, Lazzaro, Rutledge, &
97 Glimcher, 2011; Smith, Bernheim, Camerer, & Rangel, 2014; Tusche, Bode, & Haynes, 2010),
98 and that individual differences in neural activities in these regions in response to rewarding
99 stimuli are correlated with self-reported positive affect or preference for the stimuli (Bjork et al.,
100 2004; Hariri et al., 2006; Knutson, Adams, Fong, & Hommer, 2001; Knutson, Taylor, Kaufman,
101 Peterson, & Glover, 2005; Wu, Bossaerts, & Knutson, 2011). Furthermore, prior neuroimaging
102 studies have shown the involvement of these reward related regions in explicit (but not implicit)
103 self-esteem, as measured by a standardized questionnaire (i.e., trait self-esteem) (Chavez &
104 Heatherton, 2015; Frewen, Lundberg, Brimson-Theberge, & Theberge, 2013; Oikawa et al.,
105 2012) as well as momentary shift in how individuals feel about themselves (i.e., state self-
106 esteem; Will, Rutledge, Moutoussis, & Dolan, 2017). The results of a more recent study (Chavez,
107 Heatherton, & Wagner, 2017) also indicated that people's tendency to view themselves in a
108 positive manner is reflected in neural activations in the vmPFC, suggesting that, like preferences
109 for consumer goods, positive attitudes toward the self are associated with activity in reward-
110 related brain regions. In other words, neural responses in the reward-related brain regions while
111 viewing self-face is an appropriate criterion variable, because of a close theoretical fit between
112 what the self-esteem IAT scores and the neural responses should reflect (i.e., automatic
113 evaluation of the self).

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

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

122 In doing so, we employed a functional neuroimaging technique (functional magnetic
123 resonance imaging or fMRI) combined with a machine learning technique called multi-voxel
124 pattern analysis (MVPA; Haynes & Rees, 2006; Norman, Polyn, Detre, & Haxby, 2006). MVPA
125 is known to be more sensitive in detecting different psychological, cognitive, or perceptual states
126 than conventional fMRI data analysis (Izuma et al., 2017; Jimura & Poldrack, 2012; Sapountzis,
127 Schluppeck, Bowtell, & Peirce, 2010) and thus suitable for identifying potentially complex
128 associations between implicit self-esteem and neural signals in reward-related brain regions (see
129 Methods for more details). Indeed, using MVPA, a previous fMRI study (Ahn et al., 2014)
130 demonstrated that it is possible to predict individual differences in attitudes (political ideology)
131 based on brain activities. Ahn et al. (2014) found that a correlation between actual political
132 attitudes measured by a questionnaire and predicted attitudes based on MVPA was fairly high (r
133 = 0.82), suggesting that MVPA is a reliable method for identifying the relation between an
134 attitude measure and brain activities.

135 We scanned the brains of 43 individuals via fMRI while presenting them with pictures of
136 their own face (Figure 1; see Methods for power analysis). We instructed participants to carry out
137 a simple attention task while viewing pictures; we did not ask them to consider how they felt

138 about themselves. Following the fMRI scanning, each participant completed the self-esteem IAT
139 (Greenwald & Farnham, 2000) as well as two explicit self-esteem measures: (1) Rosenberg Self-
140 Esteem Scale (RSES; Rosenberg, 1965) and (2) State Self-Esteem Scale (SSES; Heatherton &
141 Polivy, 1991). By applying MVPA to the fMRI data, we were able to test whether participants'
142 level of implicit self-esteem was reliably predicted from neural signals obtained while viewing
143 their own faces. We further examined whether explicit self-esteem scores (RSES) can be
144 similarly predicted by neural signals in the reward-related brain regions, aiming to provide
145 evidence for the divergent validity of implicit versus explicit self-esteem.

146 **Method**

147 **Participants**

148 We recruited 48 women from the Neuroimaging Centre participant pool. All participants
149 were current students at the University of XXX. The literature suggests gender differences in
150 self-esteem (Bleidorn et al., 2016; Kling, Hyde, Showers, & Buswell, 1999) as well as in the
151 relationship between perceived self-face attractiveness and self-esteem (Pliner, Chaiken, & Flett,
152 1990). Thus, while passive viewing of self-face would induce neural signals related to automatic
153 evaluation of the self in both genders, the sensitivity of such responses might differ across
154 genders. Accordingly, we recruited only females in an effort to bypass such differences in this
155 first, validation study. Other inclusion criteria were: (1) ages of 18 to 28, (2) right-handedness¹,
156 (3) native command of the English language, (4) no history of neurological or psychiatric illness,
157 and (5) no metal body implants or devices. We excluded five participants from the analyses:
158 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.

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

163 **Power Analysis**

164 We estimated the effect size to be $r = 0.392$ based on a previous investigation (Ahn et al.,
165 2014). As in the present study, Ahn et al. (2014) attempted to predict individual difference in
166 social attitudes on the basis of fMRI signals. They focused on political attitudes, and reported
167 that the correlation between predicted and actual attitudes across participants ($N = 83$) was $r =$
168 0.82 . One crucial difference between Ahn et al.'s investigation and the present study is that our
169 behavioral measure (i.e., IAT) is likely to be noisier than their measure of political attitudes. We
170 estimated the difference in measurement noise based on test-retest reliability. Ahn et al. (2014)
171 reported that the test-retest reliability of political attitudes was $r = 0.952$, whereas the test-retest
172 reliability of the self-esteem IAT is $r = 0.455$; this is the average reliability of the following five
173 studies (weighted by number of participants): $r = 0.69$ (Bosson et al., 2000), $r = 0.54$ (Krause et
174 al., 2011), $r = 0.54$ (Rudolph et al., 2008, Study 1), $r = 0.52$ (Greenwald & Farnham, 2000), $r =$
175 0.39 (Rudolph et al., 2008, Study 3), and $r = 0.31$ (Gregg & Sedikides, 2010). Based on this
176 information, we estimated an effect size of $r = 0.392$ for our study. With such an effect size, a
177 sample size of $n = 39$ would achieve statistical power of $\beta = 0.2$, $\alpha = 0.05$ (one-tailed). In order to
178 account for potential data loss (e.g., due to excessive head motion in the scanner), we aimed to
179 recruit a total of 45 participants and ended up recruiting 48.

180 **Pre-Screening**

181 To ensure that our sample was characterized by a wide range of self-esteem, we emailed

182 those who expressed an interest in our fMRI study, asking them to complete an online
183 questionnaire which included the RSES. A total of 167 individuals completed the questionnaire.
184 129 of the 167 respondents were eligible for the fMRI experiment (e.g., female, 18-29 years-old,
185 right-handed, native English speakers, no history of neurological or psychiatric illness, no metal
186 in the body). The self-esteem scores of these 129 respondents were normally distributed ($range =$
187 $8-30$, $M = 19.14$, $SD = 4.66$). We invited them all for participation in the fMRI study, except for
188 most of those whose self-esteem scores hovered around the mean (16-24). Of note, the self-
189 esteem statistics (RSES score) for our final sample ($n = 43$) at the pre-screening stage were:
190 $range = 8-30$, $M = 19.88$, $SD = 5.39$.

191 **Stimuli**

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

200

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

202

203 We selected scrambled images as control stimuli, because we considered them
204 emotionally neutral. Given that we aimed to predict *individual differences* in self-esteem from

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

215 Furthermore, it could be argued that, because there are so many differences between self-
216 face and scrambled images, we cannot make strong inferences based on contrasts between these
217 conditions. There are two key differences between the present study and typical neuroimaging
218 research. First, again, the present study does not aim to identify brain regions specific to self-face
219 processing. Second, we used a machine learning technique (MVPA; see below for more detail) to
220 detect activation patterns that are associated with individual differences in the automatic
221 evaluation of the self (implicit self-esteem). Machine learning is capable of locating specific
222 patterns that are associated with a variable of interest from big (and noisy) data (Alpaydin,
223 2014). As stated above, neural signals related to individual differences in the automatic
224 evaluation of the self should be included in the contrast of the self-face versus scrambled image
225 conditions (especially in reward-related brain regions). If so, machine learning (MVPA) should
226 be able to locate specific information related to it and thus predict implicit self-esteem.

227 **Procedure**

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

235 On the second day, during fMRI scanning, participants viewed 30 blocks. These were (1)
236 self-images blocks, (2) scrambled-image control blocks, and (3) rest (i.e., a fixation cross) blocks
237 (10 blocks each). Presentation of each block lasted 12 sec. In each of the self-image and
238 scrambled-image blocks, we presented 4 different images for 2 sec each in randomized order per
239 block (inter-stimulus interval = 1 sec). Within each block, at random intervals one image
240 darkened for 300ms, which participants were instructed to detect and respond to as quickly as
241 possible with a right index finger button press. We asked participants to engage in this simple
242 task inside the scanner in order to ensure that they were paying attention to the presented images.
243 Similar low-demanding tasks have been used in studies that examined neural responses related to
244 automatic evaluations of various stimuli (Ahn et al., 2014; Cunningham et al., 2004; Izuma et al.,
245 2017; Smith et al., 2014). We recorded participants' responses within a 2 sec window post-
246 luminance change. Given that we were interested in how individual differences in implicit self-
247 esteem are related to brain activations, we fixed the order of blocks across all participants. After
248 the fMRI run (a total of 6 min), each participant underwent a different fMRI run, which is
249 unrelated to the objective of the current study (and the relevant data will not be reported here).

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

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

258 Following the IAT, we administered the explicit self-esteem measures of RSES and SSES.
259 Note that the SSES consists of three sub-scales: appearance, performance, and social. The
260 subscales assess aspects of self-esteem that are based on physical appearance, ability, and others'
261 evaluation, respectively. Finally, participants rated the attractiveness of their face (“how
262 attractive do you think your face is compared to average students on campus”) on a 7-point scale
263 (1 = *Least Attractive*, 4 = *Average*, 7 = *Most Attractive*). Upon completion, we paid participants
264 £16 and debriefed them.

265 **Behavioral Data Analysis**

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

$$272 \quad d = t[2(1-r)/n]^{1/2}$$

273

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

275 fMRI Data Acquisition

276 We used an 8 Channel head coil, GE 3T HDx Excite MRI scanner in the Neuroimaging
277 Centre to acquire whole brain fMRI data. Participants underwent a 13 second standard localizer
278 scan and 12 second ASSET calibration for parallel imaging. We also obtained high resolution T1-
279 structural scans (TE = 3 minute minimum full; TR = 7.8ms; TI = 450ms; 20° flip angle matrix =
280 256x256x176; FOV = 290x290x176; slice thickness = 1.13x1.13x1mm voxel size). Functional
281 data collection consisted of a 6 min scan, gathering 120 volumes using T2*-sensitive echo-planar
282 imaging (TE = 30ms; TR = 3000ms; 90° flip angle; matrix = 96x96; FOV = 288mm). We used
283 horizontal orientation interleaved bottom-up acquisition, with 38 4mm slices (128x128 voxels
284 per slice; 2mm voxel).

285 fMRI Data Pre-processing

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

296 Univariate fMRI Data Analysis

297 We first ran a conventional general linear model (GLM) analysis where the signal time
298 course for each participant was modeled with a GLM (Friston et al., 1995). In the GLM, we
299 modeled separately (duration = 12 sec) each of the self and scrambled-image blocks. We
300 generated regressors of interest (condition effects) using a box-car function convolved with a
301 hemodynamic-response function. We excluded regressors that were of no interest: six head
302 motion parameters (translation: x, y, and z; rotations: pitch, roll, and yaw) and high-pass filtering
303 (128 s). We created a contrast image for Self-image versus Scrambled-image for each participant,
304 and used it in subsequent MVPA analyses (see below).

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

317 **MVPA**

318 In order to predict self-esteem IAT scores from neural signals, we employed MVPA
319 (Haynes & Rees, 2006; Norman et al., 2006). In contrast to the traditional fMRI data analysis

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

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

338 We computed prediction performance using the 6-fold balanced cross-validation procedure
339 (Cohen et al., 2010; see also Kohavi, 1995); we first divided participants into 6 groups (7-8
340 participants in each group), such that these 6 groups had roughly the same means and variances
341 of self-esteem IAT scores (or RSES scores when predicting explicit self-esteem). We left out one
342 group in each cross-validation and conducted the SVR using the data from participants in all

343 other groups (training data). The SVR uses the training data to learn a weight value for each
344 voxel in a ROI, which represents the contribution of a particular voxel to predicting self-esteem
345 scores. Then, these weights are tested on participants in the left-out group (predicted IAT scores
346 for each participant in the left-outgroup is computed based on their neural signals). We repeated
347 this procedure for each group (a total of 6 times), and computed a Pearson's correlation
348 coefficient between actual IAT scores and predicted scores.

349 We tested whether brain activations in the reward-related regions predicted self-esteem IAT
350 scores. We defined the reward-related brain areas based on Neurosynth
351 (<http://www.neurosynth.org/>; Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). We used
352 an activation map from the term "Reward" (reverse inference map only), thresholded at $q < 0.01$
353 False Discovery Rate (FDR) corrected. This ROI (Figure 2a; a total of 2,696 voxels; note that we
354 used the largest cluster only) comprises brain regions that are preferentially implicated in
355 neuroimaging studies, which addressed the neural bases of reward processing² and included
356 areas involved in reward processing such as vmPFC, caudate nucleus, and midbrain (Figure 2a).
357 We also conducted the same analysis using a ROI defined by a meta-analysis (Bartra et al.,
358 2013). This meta-analysis identified a network of brain regions positively associated with
359 subjective value including bilateral striatum, vmPFC, bilateral insula, anterior cingulate cortex
360 (ACC), posterior cingulate cortex (PCC), and midbrain (brainstem). This amount to a total of
361 3,680 voxels; see Figure 3A in Bartra et al., 2013).

362 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.

363 ran MVPA using the following two ROIs. *First*, the large reward ROI (Figure 2a) included
364 medial prefrontal cortex (mPFC) regions, especially its ventral part (vmPFC). Given that mPFC
365 is known to be involved in self-processing (Denny, Kober, Wager, & Ochsner, 2012; Northoff et
366 al., 2006), which might be related to implicit or explicit self-esteem, we excluded these regions
367 from the reward ROI by applying anatomical masks (in particular, vmPFC, mPFC, and anterior
368 cingulate cortex [ACC]) using a WFU pickatlas toolbox for SPM (Maldjian, Laurienti, Kraft, &
369 Burdette, 2003). The new ROI (Figure 3a) consists of a total of 2,179 voxels. *Second*, in order to
370 limit our ROI only to regions that are even more selective to reward, we thresholded the reverse-
371 inference map obtained from Neurosynth (Figure 2a) at $z\text{-score} = 10$.³ The higher threshold
372 eliminated not only regions in the frontal cortex (e.g., vmPFC, ACC) but also other regions (e.g.,
373 putamen, thalamus, amygdala) that are relatively less selective to reward. The new ROI (Figure
374 3b) consists only of bilateral ventral striatum (nucleus accumbens) and midbrain (a total of 343
375 voxels), which are known to be the center of the reward circuit (Haber & Knutson, 2010). It is
376 well established that midbrain is rich in dopamine neurons, which encode reward-related
377 information (e.g., reward prediction error; Schultz, 2015). Similarly, ventral striatum (nucleus
378 accumbens), which is heavily interconnected with midbrain (Haber & Knutson, 2010), is known
379 to be highly sensitive (Bartra et al., 2013; Sescousse et al., 2013) and is selective to reward
380 (Ariely & Berns, 2010).

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

³ We selected $z\text{-score} = 10$, because with any $z\text{-score}$ lower than 10, there were active voxels in the frontal cortex.

385 putamen; (7) right putamen; (8) ACC; (9) left amygdala; (10) right amygdala; (11) left thalamus;
386 (12) right thalamus; and (13) midbrain. Each of the 13 reward-ROIs are known to contain
387 neurons that encode rewards or values (Komura et al., 2001; Mizuhiki, Richmond, & Shidara,
388 2012; Nishijo, Ono, & Nishino, 1988; Schultz, Apicella, & Ljungberg, 1993; for reviews, see:
389 Kolling et al., 2016; Schultz, 2015) and has been consistently activated in response to various
390 types of social and non-social rewards in human neuroimaging studies (Bartra et al., 2013;
391 Izuma, 2015; Sescousse et al., 2013). We also examined whether self-esteem scores could be
392 predicted by activation patterns in areas that were not previously implicated in reward. We
393 selected the non-reward related anatomical ROIs as follows. *First*, using Neurosynth, we
394 obtained another activation map from the term "Reward," but this time the map included both
395 reverse and forward inference maps, thresholded at $q < 0.05$ FDR corrected. This map (a total of
396 5,605 voxels) includes brain regions that were consistently (but not necessarily selectively)
397 activated in previous studies which focused on the neural bases of reward processing. *Second*, we
398 selected all anatomical structures that are not included in this broadly-defined reward-related
399 regions. These non-reward ROIs mainly include areas in parietal, temporal and occipital cortices
400 (a total of 55 non-reward ROIs; see Table 3 below for the full list of the 55 ROIs). We created all
401 of the anatomical ROIs using a WFU pickatlas toolbox for SPM (Maldjian et al., 2003).

402 Similarly, for exploratory MVPA analyses, we defined self-related brain regions using
403 Neurosynth. We used an activation map from the term "Self" (reverse inference map only),
404 thresholded at $q < 0.01$ FDR corrected. This ROI consists of two cluster (Figure 5a): (1) mPFC
405 (421 voxels), and (2) posterior cingulate cortex (PCC; 186 voxels), both of which are areas
406 previously identified in meta-analyses of fMRI studies on self-processing (Denny et al., 2012;
407 Northoff et al., 2006).

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

420 **Results**

421 **Behavioral Results**

422 Self-esteem IAT scores were significantly positive (mean IAT D score = 0.69, $t(42) =$
423 12.58, $p < 0.001$, Cohen's $d = 1.90$). Also, the self-esteem IAT was uncorrelated with the RSES (r
424 = -0.07, $p = 0.63$; a 95% confidence interval of the correlation was -0.36 to 0.24). This
425 correlation is slightly lower, but compatible with prior findings (Hofmann, Gawronski,
426 Gschwendner, Le, & Schmitt, 2005). The RSES was significantly correlated with each sub-scale
427 of the SSES (see Table 1 for all correlational results). Of note, the self-esteem IAT was related
428 neither to self-face attractiveness ratings ($r = -0.23$, $p = 0.14$) nor the appearance sub-scale of
429 SSES ($r = -0.01$, $p = 0.94$), whereas these two measures were significantly correlated with the
430 RSES ($r_s > 0.49$, $p_s < 0.001$; Table 1). Thus, any of the fMRI results reported below are unlikely

431 to be explained by participants' perceived self-face attractiveness.

432 Inside the scanner, we instructed participants to press a button when luminance of an image
433 changed. The average performance of this detection task was 96.6% for the self-image blocks
434 and 93.8% for the scrambled-image blocks, and they were not significantly different from each
435 other ($t(41) = 1.86, p = 0.07, d = 0.33$). Average reaction times were faster in the self-image
436 block (431 ms) compared to the scrambled image blocks (453 ms) ($t(41) = 2.02, p = 0.05, d =$
437 0.19), suggesting that participants' own self-faces were more attention grabbing. Importantly,
438 however, neither the self-esteem IAT ($r = -0.19, p = 0.22$) nor the RSES ($r = -0.10, p = 0.52$) was
439 related to reaction times in the self-image blocks.

440 **fMRI Results (MVPA)**

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

449

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 $p_{perm} = 0.014$.

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

451

452 Although we selected the above two ROIs based on Neurosynth term-based meta-
453 analysis (Figure 2a) and a meta-analysis of fMRI studies (Bartra et al., 2013), these regions are
454 not perfectly selective to reward. Thus, it is possible that neural signals in these ROIs and
455 implicit self-esteem were related not because these regions are involved in automatic evaluation
456 of the self, but due to other reasons like self-processing. To examine this possibility, we ran the
457 same MVPA with another ROI (Figure 3a) that does not include regions in the frontal cortex
458 such as mPFC and vmPFC, both of which are implicated in self-processing (Denny et al., 2012;
459 Northoff et al., 2006). Neural signals in the ROI predicted implicit self-esteem ($r = 0.38$, $p_{perm} =$
460 0.026). We also run the MVPA using only regions that are highly selective to reward (Figure 3b).
461 Even with this conservative test (we likely discarded at least some reward-related signals by
462 limiting our analyses to the small region), neural signals in these regions predicted implicit self-
463 esteem ($r = 0.36$, $p_{perm} = 0.036$).

464

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

466

467

468 We further tested whether the self-esteem IAT could be predicted by neural signals in each
469 of different anatomical areas, which have been implicated in reward processing. We ran the
470 MVPA with the self-esteem IAT scores within each of the 13 reward ROIs. Self-esteem IAT
471 scores were significantly predicted by neural signals in 3 out of the 13 ROIs (vmPFC, left
472 pallidum, and midbrain; Table 2). Furthermore, although prediction performances did not reach

473 the significance in the other 10 ROIs, on average, the self-esteem IAT was significantly
474 associated with activation patterns in the 13 reward ROIs (average $r = 0.24$, $t[12] = 5.42$, $p_{perm} =$
475 0.008 ; Figure 4). In contrast, neural signals in the 55 non-reward ROIs (Table 3) were, on
476 average, unrelated to the self-esteem IAT ($t[54] = 2.22$, $p_{perm} = 0.23$; Figure 4). The difference
477 between the two groups of ROIs was significant ($t[66] = 3.00$, $p_{perm} = 0.046$). These results
478 indicate that self-esteem IAT scores are related to neural signals in the reward related brain
479 regions, but not to neural signals in the non-reward related brain regions, thus further providing
480 evidence for the validity of implicit self-esteem IAT.

481

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

483

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

485 We repeated the same MVPA analyses using the explicit self-esteem (RSES) scores instead
486 of the self-esteem IAT. The large reward-related ROI (Figure 2a) was not predictive of the RSES
487 ($r = -0.08$, $p_{perm} = 0.67$). Prediction performances (correlations) using neural signals from the two
488 additional reward ROIs (Figure 3) were not significant either ($r_s < -0.03$, $p_{perm} > 0.50$).

489 Furthermore, when we applied the MVPA in each anatomical region among 13 reward-ROIs, the
490 average prediction performance was not significantly different from zero ($t[12] = 1.05$, $p_{perm} =$
491 0.61) and from the average performance of the 55 non-reward ROIs ($t[66] = 2.14$, $p_{perm} = 0.23$;
492 Tables 2 and 3), although prediction performances were significant in 3 of 13 ROIs (i.e., vmPFC,
493 right pallidum, left putamen; Table 2). Thus, explicit self-esteem was not associated with neural
494 signals in the reward related areas.

495 Furthermore, although both the self-esteem IAT and RSES were associated with at least

496 some of the reward ROIs at uncorrected $p_{perm} < 0.05$ level (Tables 2 and 3), among the 13 reward
 497 ROIs, the prediction performances were uncorrelated between the self-esteem IAT and RSES (r
 498 $= -0.37$, $p_{perm} = 0.24$). They were also uncorrelated across all 68 ROIs ($r = -0.06$, $p_{perm} = 0.91$).
 499 Moreover, the results showed that neural signals only in the vmPFC were commonly associated
 500 with both the self-esteem IAT and RSES (Table 2), indicating that neural signals in the vmPFC
 501 are linked with individual differences in both implicit and explicit self-esteem. However, when
 502 we computed a correlation between weight values of the self-esteem IAT and RSES, they were
 503 uncorrelated ($r = 0.11$, $p_{perm} = 0.21$), suggesting that the contribution of each voxel within the
 504 vmPFC to the prediction of the self-esteem IAT versus RSES differed.

505 **Exploratory MVPA Results**

506 Having provided the evidence supporting the validity of self-esteem IAT, we examined
 507 whether the self-esteem IAT (and also the RSES) is related to neural signals in other regions.⁵
 508 Particularly, given that self-esteem refers to how individuals view themselves, neural signals in
 509 regions involved in self-reference processing, namely mPFC and PCC (Denny et al., 2012;
 510 Northoff et al., 2006), may be related to the self-esteem IAT and/or the RSES. To test this
 511 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 ($-0.22 < rs < 0.08$, $p_{perm} > 0.35$).

512 voxels; Figure 5a). Interestingly, we found that neural signals in the self-related brain regions
 513 significantly predicted both the self-esteem IAT ($r = 0.50$, $p_{perm} = 0.005$; Figure 5b) and the
 514 RSES ($r = 0.39$, $p_{perm} = 0.023$; Figure 5c). We also examined whether neural signals in each of
 515 the mPFC and PCC ROIs predicted implicit and explicit self-esteem. Indeed, the self-esteem IAT
 516 was significantly predicted by neural signals in the mPFC ($r = 0.49$, $p_{perm} = 0.009$), and the PCC
 517 showed a similar trend ($r = 0.31$, $p_{perm} = 0.065$). In contrast, explicit self-esteem was not
 518 predicted by neural signals in either mPFC ($r = 0.18$, $p_{perm} = 0.18$) or PCC ($r = -0.12$, $p_{perm} =$
 519 0.67). Furthermore, although neural signals in the self-related ROI (607 voxels; Figure 5a)
 520 predicted both the self-esteem IAT and RSES, weight values of the self-esteem IAT and RSES
 521 were uncorrelated with each other, indicating that they are represented differently in these
 522 regions ($r = -0.06$, $p_{perm} = 0.67$).⁶

523

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

525

526 Another possibility is that implicit (and explicit) self-esteem may modulate how
 527 individuals view their faces, and thus may be related to neural signals in regions involved in face
 528 processing such as fusiform gyrus. Consistent with this possibility, an fMRI study has
 529 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.

530 correlated with implicit prejudice (i.e., race IAT scores; Cunningham et al., 2004). Further, more
531 recent MVPA studies indicate that neural signals in fusiform face area (FFA) in response to faces
532 are modulated depending on implicit attitudes (Brosch, Bar-David, & Phelps, 2013) or
533 stereotypes (Stolier & Freeman, 2016). However, our results showed that activations in both left
534 and right fusiform gyrus were unassociated with the self-esteem IAT (left $r = 0.21$, $p_{perm} = 0.17$;
535 right $r = 0.26$, $p_{perm} = 0.12$; Table 3), although both correlations were in a positive direction. The
536 RSES was also unassociated with activations in fusiform gyrus (left $r = -0.46$, $p_{perm} = 0.99$; right
537 $r = -0.09$, $p_{perm} = 0.65$).⁷

538 **fMRI Results (Univariate Analysis)**

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

545 **Discussion**

546 We aimed to provide unique evidence for the validity of an implicit self-esteem measure
547 using neuroimaging combined with a machine learning technique, MVPA. Our findings indicate
548 that implicit self-esteem, as measured by the IAT, is associated with neural activation patterns
549 automatically evoked by passive viewing of self-face in the reward-related regions (Figures 2a,
550 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.

551 non-reward related areas (Table 3 and Figure 4). Thus, although in prior research (Falk & Heine,
552 2015) implicit self-esteem measures were found to be unrelated to personality or attitude
553 measures (i.e., had low convergent and predictive validity), in our study self-esteem IAT scores
554 were robustly associated with (i.e., predictive of) neural signals in a way that is consistent with
555 the conceptualization of implicit self-esteem as an automatic attitude toward the self (Greenwald
556 & Banaji, 1995; Sedikides & Gregg, 2003). The literature has indicated that attractive faces
557 activate reward-related brain areas (Cloutier, Heatherton, Whalen, & Kelley, 2008; O'Doherty et
558 al., 2003), and that face attractiveness can be decoded from neural signals in vmPFC (Pegors,
559 Kable, Chatterjee, & Epstein, 2015). However, given that IAT scores were unrelated to both
560 perceived self-face attractiveness and the SSES sub-scale of appearance (while both being
561 significantly related to explicit self-esteem scores, i.e., RSES), our results are unlikely to be
562 mediated by individual difference in perceived self-face attractiveness.

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

571 Similarly, as stated earlier, the low predictive validity of implicit self-esteem measures may
572 be due to biases in selecting criterion variables, which is likely due to lack of clear understanding
573 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,

597 2006), which postulates that implicit and explicit evaluations are the outcomes of two distinct
598 processes: (1) associative, and (2) propositional. The APE model states that, although implicit
599 evaluations depend on associative processes (i.e., automatically activated associations), explicit
600 evaluations depend on activated associations (associative processes) and their validation
601 according to cognitive consistency principles (propositional processes). It is, of course, rather
602 simplistic to regard the associative and propositional processes of the APE model as mapping
603 directly onto the mPFC and PCC, respectively. Yet, it is possible that explicit self-esteem is
604 determined by a similar interaction process between two (unspecified) distinct processes.

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

620 while participants are engaging in explicit evaluations of self (e.g., self-reference task), can
621 predict individual differences in explicit self-esteem and differences/similarities in how implicit
622 and explicit self-esteem are represented in these regions.

623 We based the study's design on findings that activity in the reward related brain regions as
624 a response to an object reflects participants' preference for that object (Izuma et al., 2017;
625 Lebreton et al., 2009; Levy et al., 2011; Smith et al., 2014; Tusche et al., 2010). One might argue,
626 however, that evidence could have been stronger, if we demonstrated that a decoder of
627 preference for non-social reward objects (e.g., food) could predict implicit self-esteem (i.e., a
628 more direct link between activity in the reward related areas and neural signals as a response to
629 self-face). Here, we would first train the prediction model on responses to a food, then apply this
630 model to neural responses to one's own face, and finally test if it can predict self-esteem IAT
631 scores. Although such a demonstration would have been ideal, this proposal would rely on the
632 assumption that preferences for non-social objects and attitudes toward the self are represented in
633 a similar manner in the brain. Such an assumption is empirically unsupported. Previous
634 neurophysiological studies with monkeys and rats established that largely distinct populations of
635 striatal neurons encode reward values of different types of reward (e.g., juice vs. drug rewards;
636 Bowman, Aigner, & Richmond, 1996; Carelli, Ijames, & Crumling, 2000; Carelli &
637 Wondolowski, 2003; Robinson & Carelli, 2008). A recent MVPA study also indicated that,
638 although there may exist a population of neurons that encode both social and non-social rewards,
639 these two types of rewards are processed in largely distinct neural circuits (Wake & Izuma,
640 2017).

641 We recruited only young female individuals in Western culture. It is interesting and
642 important to test whether the findings can be replicated in males or individuals from different

643 cultures. In addition to testing the validity of the self-esteem IAT, our study also afforded a novel
644 insight into what implicit self-esteem (as measured by the IAT) is by demonstrating an
645 association between neural signals in self-processing regions (i.e., mPFC and PCC) and implicit
646 (and explicit) self-esteem. Prior research (Kitayama & Uchida, 2003; Yamaguchi et al., 2007)
647 showed that, whereas people in Western countries tend to have higher explicit self-esteem than
648 those in East-Asian countries, both cultures manifest the same level of implicit self-esteem (for a
649 review, see: Sedikides, Gaertner, & Cai, 2015). Future empirical efforts could be directed toward
650 addressing similarities/differences between Western and Eastern cultures in terms of neural
651 representations of implicit and explicit self-esteem.

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

662

663 **Acknowledgements**

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

667 **References**

- 668 Ahn, W. Y., Kishida, K. T., Gu, X., Lohrenz, T., Harvey, A., Alford, J. R., . . . Montague, P. R.
669 (2014). Nonpolitical images evoke neural predictors of political ideology. *Current*
670 *Biology*, 24(22), 2693–2699.
- 671 Alpaydin, E. (Ed.) (2014). *Introduction to machine learning (3rd ed)*. Cambridge, MA: The MIT
672 Press.
- 673 Amodio, D. M. (2010). Can neuroscience advance social psychological theory? Social
674 neuroscience for the behavioral social psychologist. *Social Cognition*, 28(6), 695-716.
- 675 Ariely, D., & Berns, G. S. (2010). Neuromarketing: the hope and hype of neuroimaging in
676 business. *Nature Reviews Neuroscience*, 11(4), 284-292. doi:10.1038/nrn2795
- 677 Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude
678 measures. *Behavioral Research Methods*, 46(3), 668-688. doi:10.3758/s13428-013-0410-
679 6
- 680 Bartra, O., McGuire, J. T., & Kable, J. W. (2013). The valuation system: a coordinate-based
681 meta-analysis of BOLD fMRI experiments examining neural correlates of subjective
682 value. *Neuroimage*, 76, 412-427. doi:10.1016/j.neuroimage.2013.02.063
- 683 Bjork, J. M., Knutson, B., Fong, G. W., Caggiano, D. M., Bennett, S. M., & Hommer, D. W.
684 (2004). Incentive-elicited brain activation in adolescents: similarities and differences
685 from young adults. *Journal of Neuroscience*, 24(8), 1793-1802.
686 doi:10.1523/JNEUROSCI.4862-03.2004
- 687 Blanton, H., Jaccard, J., Christie, C., & Gonzales, P. M. (2007). Plausible assumptions,
688 questionable assumptions and post hoc rationalizations: Will the real IAT, please stand
689 up? *Journal of Experimental Social Psychology*, 43(3), 399-409.

690 doi:10.1016/j.jesp.2006.10.019

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

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

697 Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of
698 implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality*
699 *and Social Psychology, 79*(4), 631-643. doi:10.1037//0022-3514.79.4.631

700 Bowman, E. M., Aigner, T. G., & Richmond, B. J. (1996). Neural signals in the monkey ventral
701 striatum related to motivation for juice and cocaine rewards. *Journal of Neurophysiology,*
702 *75*(3), 1061-1073.

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

706 Buhrmester, M. D., Blanton, H., & Swann, W. B. (2011). Implicit self-esteem: Nature,
707 measurement, and a new way forward. *Journal of Personality and Social Psychology,*
708 *100*(2), 365-385. doi:10.1037/A0021341

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

712 Carelli, R. M., & Wondolowski, J. (2003). Selective encoding of cocaine versus natural rewards

- 713 by nucleus accumbens neurons is not related to chronic drug exposure. *Journal of*
714 *Neuroscience*, 23(35), 11214-11223.
- 715 Chavez, R. S., & Heatherton, T. F. (2015). Multimodal frontostriatal connectivity underlies
716 individual differences in self-esteem. *Social Cognitive and Affective Neuroscience*, 10(3),
717 364-370. doi:10.1093/scan/nsu063
- 718 Chavez, R. S., Heatherton, T. F., & Wagner, D. D. (2017). Neural population decoding reveals
719 the intrinsic positivity of the self. *Cerebral Cortex*, 27(11), 5222-5229.
720 doi:10.1093/cercor/bhw302
- 721 Cloutier, J., Heatherton, T. F., Whalen, P. J., & Kelley, W. M. (2008). Are attractive people
722 rewarding? Sex differences in the neural substrates of facial attractiveness. *Journal of*
723 *Cognitive Neuroscience*, 20(6), 941-951. doi:10.1162/jocn.2008.20062
- 724 Cohen, J. R., Asarnow, R. F., Sabb, F. W., Bilder, R. M., Bookheimer, S. Y., Knowlton, B. J., &
725 Poldrack, R. A. (2010). Decoding developmental differences and individual variability in
726 response inhibition through predictive analyses across individuals. *Frontiers in Human*
727 *Neuroscience*, 4, 47. doi:10.3389/fnhum.2010.00047
- 728 Cunningham, W. A., Johnson, M. K., Raye, C. L., Chris Gatenby, J., Gore, J. C., & Banaji, M. R.
729 (2004). Separable neural components in the processing of black and white faces.
730 *Psychological Science*, 15(12), 806-813. doi:10.1111/j.0956-7976.2004.00760.x
- 731 Cvencek, D., Greenwald, A. G., & Meltzoff, A. N. (2016). Implicit measures for preschool
732 children confirm self-esteem's role in maintaining a balanced identity. *Journal of*
733 *Experimental Social Psychology*, 62, 50-57. doi:10.1016/j.jesp.2015.09.015
- 734 Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional
735 neuroimaging studies of self- and other judgments reveals a spatial gradient for

- 736 mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 24(8), 1742-
737 1752. doi:10.1162/jocn_a_00233
- 738 Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector
739 regression machines. *Advances in Neural Information Processing Systems* 9(9), 155-161.
- 740 Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments
741 with matched groups or repeated measures designs. *Psychological Methods*, 1(2), 170-
742 177. doi:10.1037/1082-989x.1.2.170
- 743 Falk, C. F., & Heine, S. J. (2015). What is implicit self-esteem, and does it vary across cultures?
744 *Personality and Social Psychology Review*, 19(2), 177-198.
745 doi:10.1177/1088868314544693
- 746 Falk, C. F., Heine, S. J., Takemura, K., Zhang, C. X., & Hsu, C. W. (2015). Are implicit self-
747 esteem measures valid for assessing individual and cultural differences? *Journal of*
748 *Personality*, 83(1), 56-68. doi:10.1111/jopy.12082
- 749 Frewen, P. A., Lundberg, E., Brimson-Theberge, M., & Theberge, J. (2013). Neuroimaging self-
750 esteem: a fMRI study of individual differences in women. *Social Cognitive and Affective*
751 *Neuroscience*, 8(5), 546-555. doi:10.1093/scan/nss032
- 752 Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., & Frackowiak, R. S. J.
753 (1995). Statistical parametric maps in functional imaging: A general linear approach.
754 *Human Brain Mapping*, 2, 189-210. doi:10.1002/hbm.460020402
- 755 Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in
756 evaluation: An integrative review of implicit and explicit attitude change. *Psychological*
757 *Bulletin*, 132(5), 692-731. doi:10.1037/0033-2909.132.5.692
- 758 Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and

- 759 stereotypes. *Psychological Review*, 102(1), 4-27. doi:10.1037//0033-295x.102.1.4
- 760 Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., & Mellott, D. S.
761 (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept.
762 *Psychological Review*, 109(1), 3-25. doi:10.1037//0033-295x.109.1.3
- 763 Greenwald, A. G., & Farnham, S. D. (2000). Using the implicit association test to measure self-
764 esteem and self-concept. *Journal of Personality and Social Psychology*, 79(6), 1022-
765 1038. doi:10.1037//0022-3514.79.6.1022
- 766 Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences
767 in implicit cognition: The implicit association test. *Journal of Personality and Social*
768 *Psychology*, 74(6), 1464-1480. doi:10.1037/0022-3514.74.6.1464
- 769 Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit
770 association test: I. An improved scoring algorithm. *Journal of Personality and Social*
771 *Psychology*, 85(2), 197-216. doi:10.1037/0022-3514.85.2.197
- 772 Gregg, A. P., Mahadevan, N., & Sedikides, C. (2017). The SPOT effect: People spontaneously
773 prefer their own theories. *The Quarterly Journal of Experimental Psychology*, 70(6), 996-
774 1010. doi:10.1080/17470218.2015.1099162
- 775 Gregg, A. P., & Sedikides, C. (2010). Narcissistic fragility: Rethinking its links to explicit and
776 implicit self-esteem. *Self and Identity*, 9(2), 142-161. doi:10.1080/15298860902815451
- 777 Haber, S. N., & Knutson, B. (2010). The reward circuit: Linking primate anatomy and human
778 imaging. *Neuropsychopharmacology*, 35(1), 4-26. doi:10.1038/npp.2009.129
- 779 Hariri, A. R., Brown, S. M., Williamson, D. E., Flory, J. D., de Wit, H., & Manuck, S. B. (2006).
780 Preference for immediate over delayed rewards is associated with magnitude of ventral
781 striatal activity. *Journal of Neuroscience*, 26(51), 13213-13217.

- 782 doi:10.1523/JNEUROSCI.3446-06.2006
- 783 Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural
784 basis of episodic memory. *Journal of Neuroscience*, 27(52), 14365-14374.
785 doi:10.1523/JNEUROSCI.4549-07.2007
- 786 Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature*
787 *Reviews Neuroscience*, 7(7), 523-534. doi:10.1038/nrn1931
- 788 Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state
789 self-esteem. *Journal of Personality and Social Psychology*, 60(6), 895-910.
790 doi:10.1037//0022-3514.60.6.895
- 791 Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis
792 on the correlation between the implicit association test and explicit self-report measures.
793 *Personality and Social Psychology Bulletin*, 31(10), 1369-1385.
794 doi:10.1177/0146167205275613
- 795 Honekopp, J. (2006). Once more: is beauty in the eye of the beholder? Relative contributions of
796 private and shared taste to judgments of facial attractiveness. *Journal of Experimental*
797 *Psychology: Human Perception and Performance*, 32(2), 199-209. doi:10.1037/0096-
798 1523.32.2.199
- 799 Izuma, K. (2015). Social reward. In A. W. Toga (Ed.), *Brain mapping: An encyclopedic reference*
800 (Vol. 3, pp. 21-23). Oxford, UK: Elsevier.
- 801 Izuma, K., Shibata, K., Matsumoto, K., & Adolphs, R. (2017). Neural predictors of evaluative
802 attitudes towards celebrities. *Social Cognitive and Affective Neuroscience*, 12, 382-390.
803 doi:10.1093/scan/nsw135
- 804 Jimura, K., & Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel

- 805 pattern information tell complementary stories. *Neuropsychologia*, *50*(4), 544-552.
806 doi:10.1016/j.neuropsychologia.2011.11.007
- 807 Jordan, C. H., Logel, C., Spencer, S. J., Zanna, M. P., & Whitfield, M. L. (2009). The
808 heterogeneity of self-esteem. In R. E. Petty, R. H. Fazio, & P. Brinol (Eds.), *Attitudes:
809 Insight from the new implicit measures*. New York, NY: Psychology Press.
- 810 Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during
811 intertemporal choice. *Nature Neuroscience*, *10*(12), 1625-1633. doi:10.1038/nn2007
- 812 Kaplan, J. T., Aziz-Zadeh, L., Uddin, L. Q., & Iacoboni, M. (2008). The self across the senses: an
813 fMRI study of self-face and self-voice recognition. *Social Cognitive Affective
814 Neuroscience*, *3*(3), 218-223. doi:10.1093/scan/nsn014
- 815 Kitayama, S., & Uchida, Y. (2003). Explicit self-criticism and implicit self-regard: Evaluating
816 self and friend in two cultures. *Journal of Experimental Social Psychology*, *39*(5), 476-
817 482. doi:10.1016/S0022-1031(03)00026-X
- 818 Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-
819 esteem: A meta-analysis. *Psychological Bulletin*, *125*(4), 470-500. doi:10.1037//0033-
820 2909.125.4.470
- 821 Knutson, B., Adams, C. M., Fong, G. W., & Hommer, D. (2001). Anticipation of increasing
822 monetary reward selectively recruits nucleus accumbens. *Journal of Neuroscience*,
823 *21*(16), RC159.
- 824 Knutson, B., Fong, G. W., Bennett, S. M., Adams, C. M., & Hommer, D. (2003). A region of
825 mesial prefrontal cortex tracks monetarily rewarding outcomes: characterization with
826 rapid event-related fMRI. *Neuroimage*, *18*(2), 263-272.
- 827 Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural

- 828 representation of expected value. *Journal of Neuroscience*, 25(19), 4806-4812.
829 doi:10.1523/JNEUROSCI.0642-05.2005
- 830 Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model*
831 *selection*. Paper presented at the International Joint Conference on Artificial Intelligence,
832 San Fransisco, CA.
- 833 Kolling, N., Behrens, T., Wittmann, M. K., & Rushworth, M. (2016). Multiple signals in anterior
834 cingulate cortex. *Current Opinion in Neurobiology*, 37, 36-43.
835 doi:10.1016/j.conb.2015.12.007
- 836 Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., & Ono, T. (2001). Retrospective and
837 prospective coding for predicted reward in the sensory thalamus. *Nature*, 412(6846), 546-
838 549. doi:10.1038/35087595
- 839 Krause, S., Back, M. D., Egloff, B., & Schmukle, S. C. (2011). Reliability of Implicit Self-
840 esteem Measures Revisited. *European Journal of Personality*, 25(3), 239-251.
841 doi:10.1002/Per.792
- 842 Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An automatic
843 valuation system in the human brain: evidence from functional neuroimaging. *Neuron*,
844 64(3), 431-439. doi:10.1016/j.neuron.2009.09.040
- 845 Levy, I., Lazzaro, S. C., Rutledge, R. B., & Glimcher, P. W. (2011). Choice from non-choice:
846 predicting consumer preferences from blood oxygenation level-dependent signals
847 obtained during passive viewing. *Journal of Neuroscience*, 31(1), 118-125.
848 doi:10.1523/JNEUROSCI.3214-10.2011
- 849 Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for
850 neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets.

- 851 *Neuroimage*, 19(3), 1233-1239.
- 852 Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: a meta-
853 analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience*,
854 8(3), 285-299. doi:10.1093/scan/nsr090
- 855 Mizuhiki, T., Richmond, B. J., & Shidara, M. (2012). Encoding of reward expectation by
856 monkey anterior insular neurons. *Journal of Neurophysiology*, 107(11), 2996-3007.
857 doi:10.1152/jn.00282.2011
- 858 Nishijo, H., Ono, T., & Nishino, H. (1988). Single neuron responses in amygdala of alert monkey
859 during complex sensory stimulation with affective significance. *Journal of Neuroscience*,
860 8(10), 3570-3583.
- 861 Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-
862 voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9), 424-430.
863 doi:10.1016/j.tics.2006.07.005
- 864 Northoff, G., Heinzel, A., de Greck, M., Bormpohl, F., Dobrowolny, H., & Panksepp, J. (2006).
865 Self-referential processing in our brain: a meta-analysis of imaging studies on the self.
866 *Neuroimage*, 31(1), 440-457.
- 867 Nuttin, J. M., Jr. (1985). Narcissism beyond Gestalt and awareness: the name letter effect.
868 *European Journal of Social Psychology*, 15, 353-361.
- 869 O'Doherty, J., Winston, J., Critchley, H., Perrett, D., Burt, D. M., & Dolan, R. J. (2003). Beauty
870 in a smile: the role of medial orbitofrontal cortex in facial attractiveness.
871 *Neuropsychologia*, 41(2), 147-155.
- 872 Oikawa, H., Sugiura, M., Sekiguchi, A., Tsukiura, T., Miyauchi, C. M., Hashimoto, T., . . .
873 Kawashima, R. (2012). Self-face evaluation and self-esteem in young females: an fMRI

- 874 study using contrast effect. *Neuroimage*, 59(4), 3668-3676.
875 doi:10.1016/j.neuroimage.2011.10.098
- 876 Op de Beeck, H. P. (2010). Against hyperacuity in brain reading: Spatial smoothing does not hurt
877 multivariate fMRI analyses? *Neuroimage*, 49(3), 1943-1948.
878 doi:10.1016/J.Neuroimage.2009.02.047
- 879 Pegors, T. K., Kable, J. W., Chatterjee, A., & Epstein, R. A. (2015). Common and unique
880 representations in pFC for face and place attractiveness. *Journal of Cognitive*
881 *Neuroscience*, 27(5), 959-973. doi:10.1162/jocn_a_00777
- 882 Peirce, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience*
883 *Methods*, 162(1-2), 8-13. doi:10.1016/j.jneumeth.2006.11.017
- 884 Pliner, P., Chaiken, S., & Flett, G. L. (1990). Gender differences in concern with body weight
885 and physical appearance over the life-span. *Personality and Social Psychology Bulletin*,
886 16(2), 263-273. doi:10.1177/0146167290162007
- 887 Robinson, D. L., & Carelli, R. M. (2008). Distinct subsets of nucleus accumbens neurons encode
888 operant responding for ethanol versus water. *European Journal of Neuroscience*, 28(9),
889 1887-1894. doi:10.1111/j.1460-9568.2008.06464.x
- 890 Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal
891 information. *Journal of Personality and Social Psychology*, 35(9), 677-688.
892 doi:10.1037/0022-3514.35.9.677
- 893 Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton
894 University Press.
- 895 Rudolph, A., Schroder-Abe, M., Schutz, A., Gregg, A. P., & Sedikides, C. (2008). Through a
896 glass, less darkly? Reassessing convergent and discriminant validity in measures of

897 implicit self-esteem. *European Journal of Psychological Assessment*, 24(4), 273-281.

898 doi:10.1027/1015-5759.24.4.273

899 Sapountzis, P., Schluppeck, D., Bowtell, R., & Peirce, J. W. (2010). A comparison of fMRI

900 adaptation and multivariate pattern classification analysis in visual cortex. *Neuroimage*,

901 49(2), 1632-1640. doi:10.1016/j.neuroimage.2009.09.066

902 Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological*

903 *Review*, 95(3), 853-951. doi:10.1152/physrev.00023.2014

904 Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to

905 reward and conditioned stimuli during successive steps of learning a delayed response

906 task. *Journal of Neuroscience*, 13(3), 900-913.

907 Sedikides, C., Gaertner, L., & Cai, H. (2015). On the panculturality of self-enhancement and

908 self-protection motivation: The case for the universality of self-esteem. *Advances in*

909 *Motivation Science*, 2, 185-241. doi:10.1016/bs.adms.2015.04.002

910 Sedikides, C., & Gregg, A. P. (2003). Portraits of the self. In M. A. Hogg & J. Cooper (Eds.), *Sage*

911 *handbook of social psychology* (pp. 110-138). London, UK: Sage Publications.

912 Sescousse, G., Caldu, X., Segura, B., & Dreher, J. C. (2013). Processing of primary and

913 secondary rewards: a quantitative meta-analysis and review of human functional

914 neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, 37(4), 681-696.

915 doi:10.1016/j.neubiorev.2013.02.002

916 Shibata, K., Watanabe, T., Kawato, M., & Sasaki, Y. (2016). Differential activation patterns in

917 the same brain region led to opposite emotional states. *PLoS Biology*, 14(9), e1002546.

918 doi:10.1371/journal.pbio.1002546

919 Smith, A., Bernheim, B. D., Camerer, C. F., & Rangel, A. (2014). Neural activity reveals

- 920 preferences without choices. *American Economic Journal-Microeconomics*, 6(2), 1-36.
921 doi:10.1257/Mic.6.2.1
- 922 Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent
923 intersection of social categories. *Nature Neuroscience*, 19(6), 795-797.
924 doi:10.1038/nn.4296
- 925 Sugiura, M., Kawashima, R., Nakamura, K., Okada, K., Kato, T., Nakamura, A., . . . Fukuda, H.
926 (2000). Passive and active recognition of one's own face. *Neuroimage*, 11(1), 36-48.
927 doi:10.1006/nimg.1999.0519
- 928 Toga, A. W., & Thompson, P. M. (2003). Mapping brain asymmetry. *Nature Reviews*
929 *Neuroscience*, 4(1), 37-48. doi:10.1038/nrn1009
- 930 Tusche, A., Bode, S., & Haynes, J. D. (2010). Neural responses to unattended products predict
931 later consumer choices. *Journal of Neuroscience*, 30(23), 8024-8031.
932 doi:10.1523/JNEUROSCI.0064-10.2010
- 933 Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the
934 human striatum. *Social Cognitive and Affective Neuroscience*, 12(10), 1558-1564.
935 doi:10.1093/scan/nsx092.
- 936 Will, G., Rutledge, R. B., Moutoussis, M., & Dolan, R. J. (2017). Neural and computational
937 processes underlying dynamic changes in self-esteem. *eLife*, 6, e28098.
938 doi:10.7554/eLife.28098
- 939 Wu, C. C., Bossaerts, P., & Knutson, B. (2011). The affective impact of financial skewness on
940 neural activity and choice. *PLoS One*, 6(2), e16838. doi:10.1371/journal.pone.0016838
- 941 Yamaguchi, S., Greenwald, A. G., Banaji, M. R., Murakami, F., Chen, D., Shiomura, K., . . .
942 Krendl, A. (2007). Apparent universality of positive implicit self-esteem. *Psychological*

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

944 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale

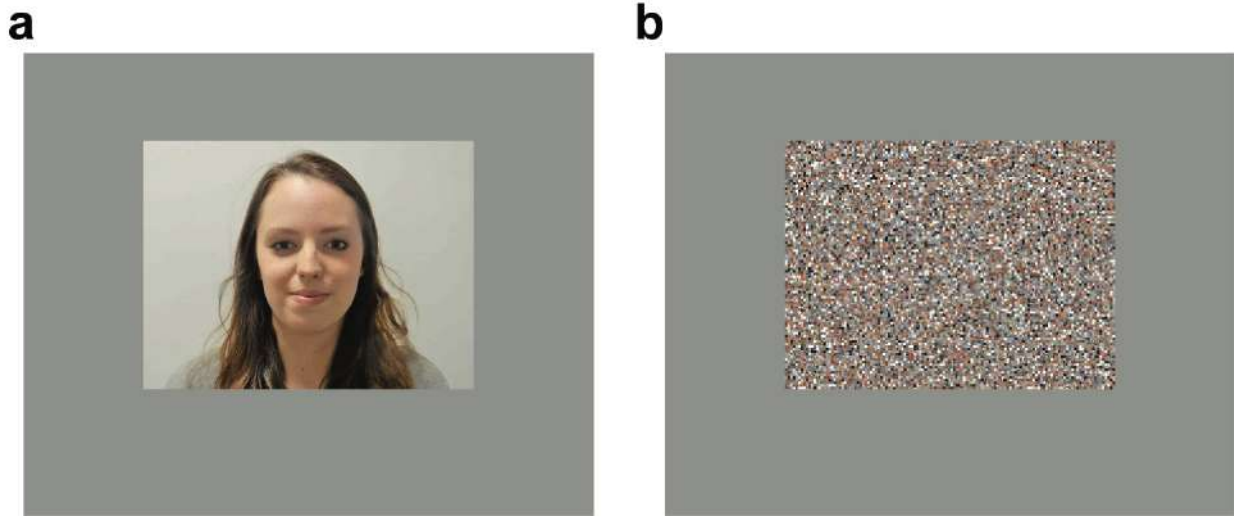
945 automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8), 665-

946 U695. doi:10.1038/Nmeth.1635

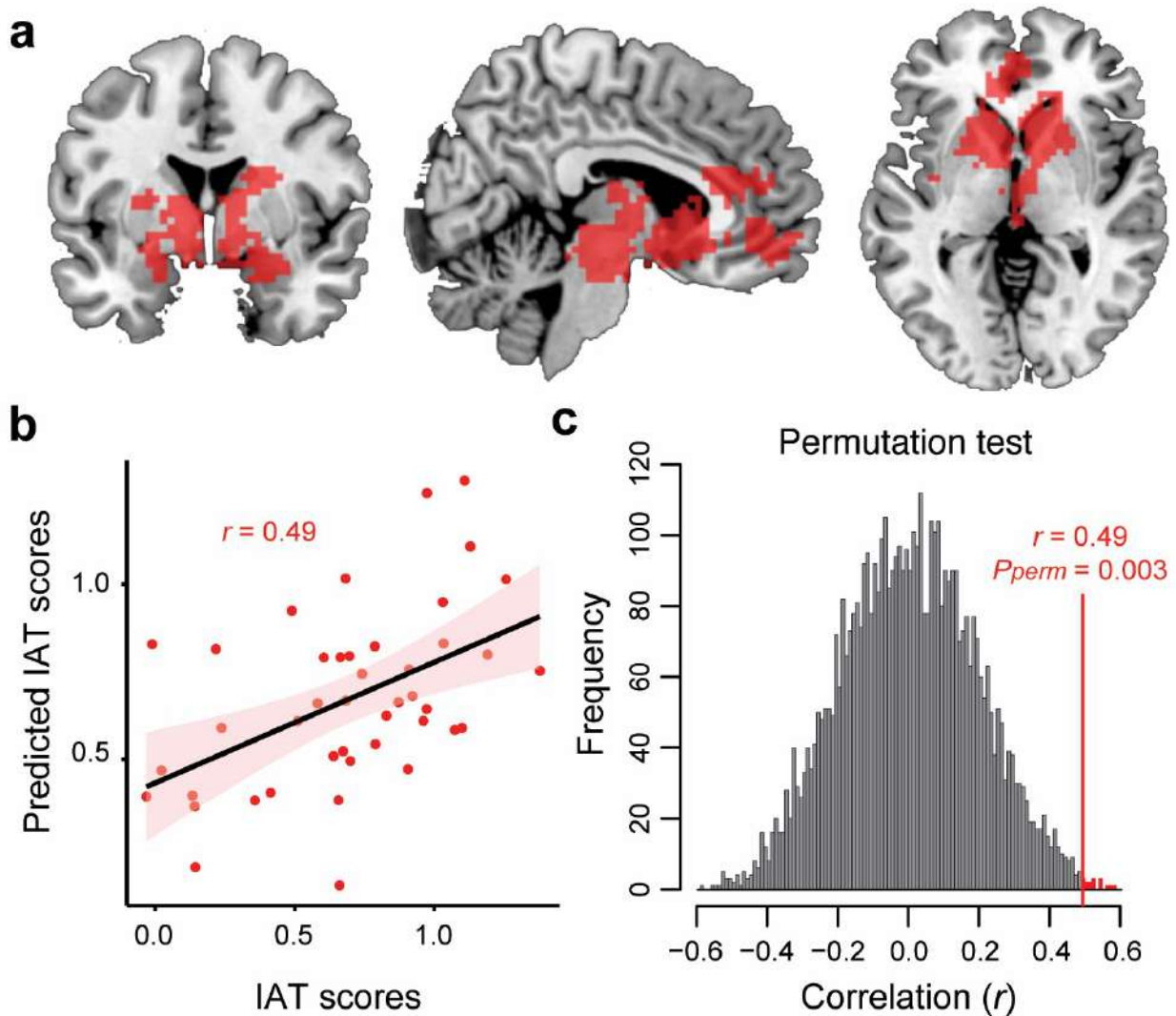
947

948

949 **Figures**



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



954

955 **Figure 2.** (a). A large reward-related ROI defined using Neurosynth (a total of 2,696 voxels).956 Left: coronal view ($y = 0$). Middle: sagittal view ($x = 6$). Right: Axial view ($z = 0$). (b). A

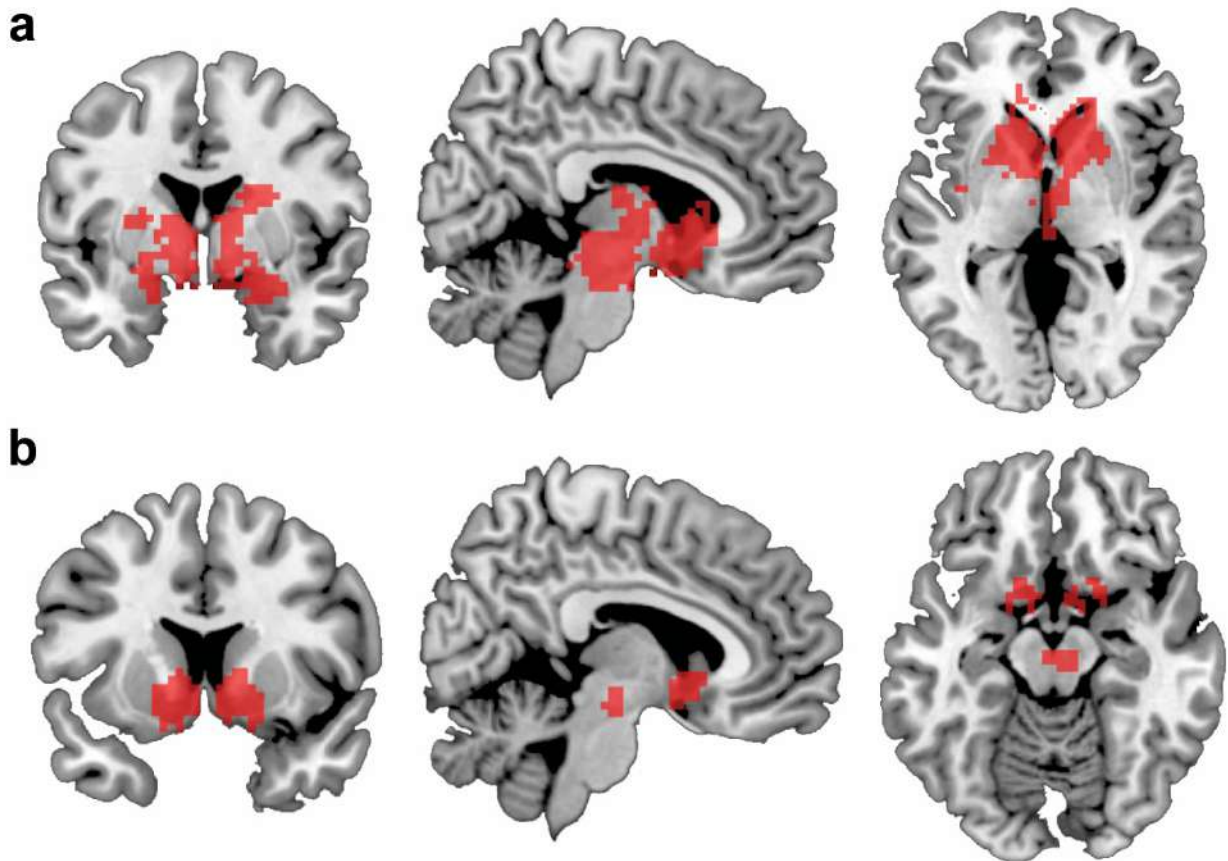
957 correlation between participants' self-esteem IAT scores and predicted scores based on neural

958 signals in the ROI. (c). A histogram showing the distributions of correlation coefficients between

959 actual and predicted IAT scores with randomly permuted data (5,000 times). The correlation

960 with actual data was significant at $p_{perm} = 0.003$.

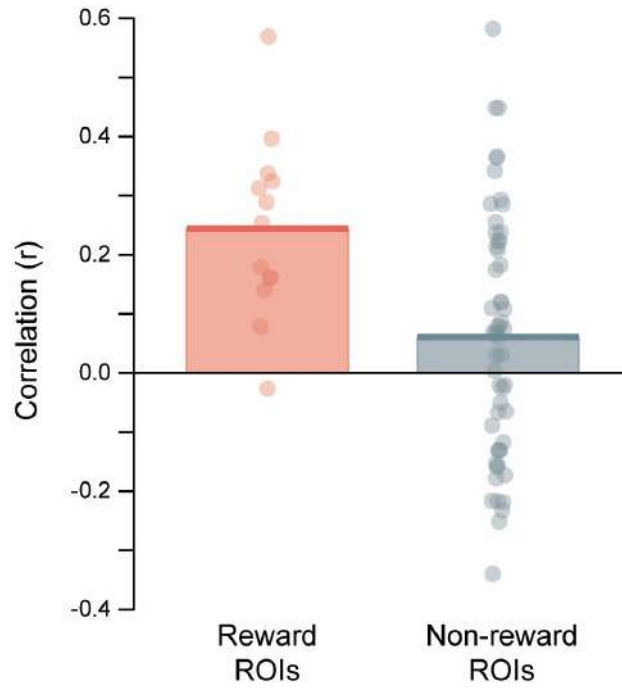
961



962

963 **Figure 3.** Two Additional Reward ROIs. (a) Anatomical structures in the frontal cortex (i.e.,
 964 mPFC, vmPFC, ACC) were removed from the large reward ROI (Figure 2a). There are a total of
 965 2,179 voxels. Left: coronal view ($y = 0$). Middle: sagittal view ($x = 6$). Right: Axial view ($z = 0$).
 966 (b) Regions highly selective to reward obtained from Neurosynth (a term-based meta-analysis
 967 with the term "Reward" and thresholded at z -score = 10). The ROI consists of bilateral ventral
 968 striatum (nucleus accumbens) and midbrain (a total of 343 voxels). Left: coronal view ($y = 10$).
 969 Middle: sagittal view ($x = 6$). Right: Axial view ($z = -14$).

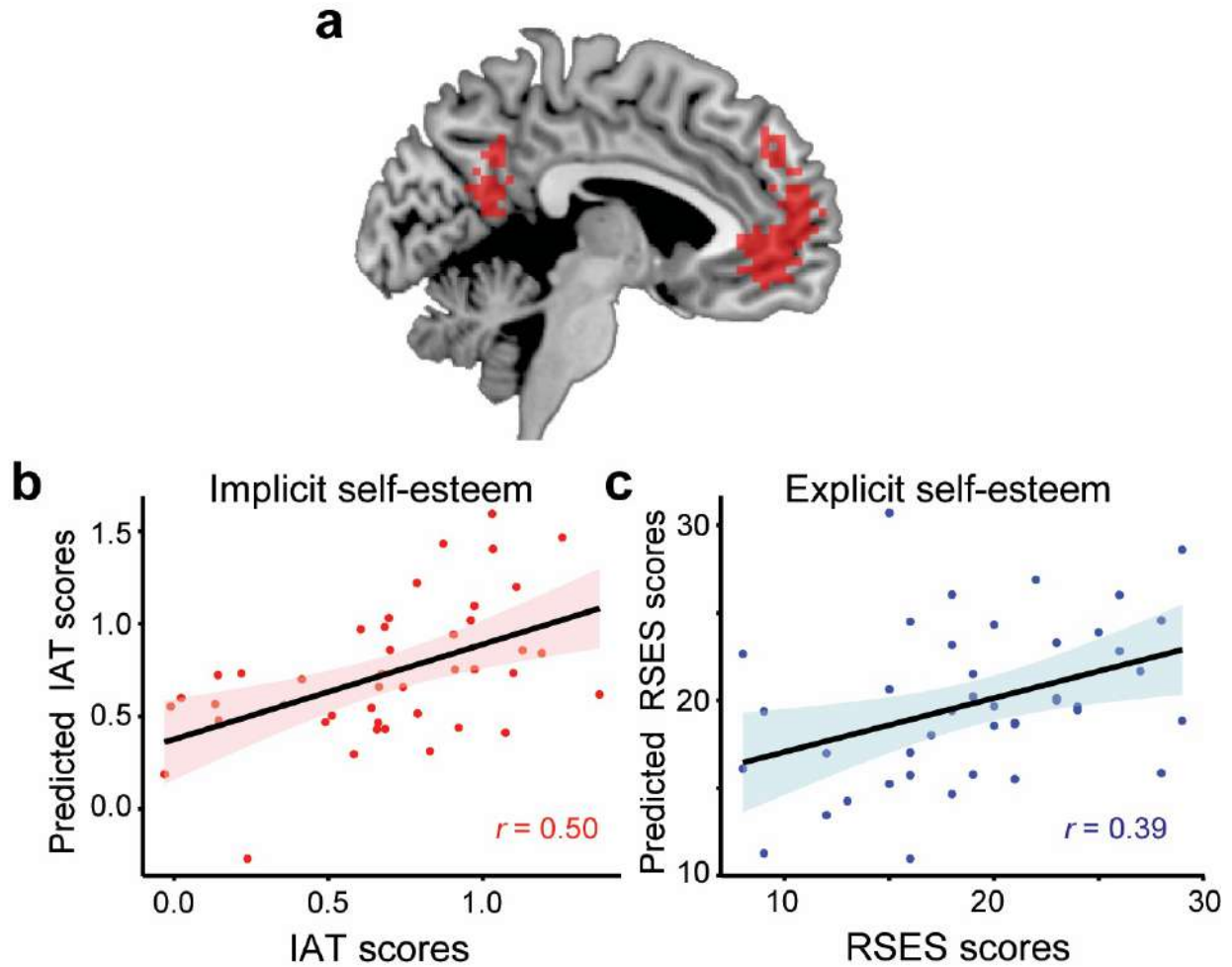
970



971

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

976



977

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