

<sup>1</sup> Positivity bias in perceptual matching may reflect a spontaneous self-referential processing

<sup>2</sup> Hu Chuan-Peng<sup>1,2</sup>, Kaiping Peng<sup>3</sup>, & Jie Sui<sup>3,4</sup>

<sup>3</sup> <sup>1</sup> TBA

<sup>4</sup> <sup>2</sup> Leibniz Institute for Resilience Research, 55131 Mainz, Germany

<sup>5</sup> <sup>3</sup> Tsinghua University, 100084 Beijing, China

<sup>6</sup> <sup>4</sup> University of Aberdeen, Aberdeen, Scotland

<sup>7</sup> Author Note

<sup>8</sup> Hu Chuan-Peng, Leibniz Institute for Resilience Research (LIR). Kaiping Peng,

<sup>9</sup> Department of Psychology, Tsinghua University, 100084 Beijing, China. Jie Sui, School of

<sup>10</sup> Psychology, University of Aberdeen, Aberdeen, Scotland.

<sup>11</sup> Authors contriubtion: HCP, JS, & KP design the study, HCP collected the data,

<sup>12</sup> HCP analyzed the data and drafted the manuscript. KP & JS supported this project.

<sup>13</sup> Correspondence concerning this article should be addressed to Hu Chuan-Peng,

<sup>14</sup> Langenbeckstr. 1, Neuroimaging Center, University Medical Center Mainz, 55131 Mainz,

<sup>15</sup> Germany. E-mail: hcp4715@gmail.com

16

## Abstract

17 To navigate in a complex social world, individual has learnt to prioritize valuable  
18 information. Previous studies suggested the moral related stimuli was prioritized  
19 (Anderson, Siegel, Bliss-Moreau, & Barrett, 2011; Gantman & Van Bavel, 2014). Using  
20 social associative learning paradigm (self-tagging paradigm), we found that when geometric  
21 shapes, without soical meaning, were associated with different moral valence (morally  
22 good, neutral, or bad), the shapes that associated with positive moral valence were  
23 prioritized in a perceptual matching task. This patterns of results were robust across  
24 different procedures. Further, we tested whether this positive effect was modulated by  
25 self-relevance by manipulating the self-referential explicitly and found that this moral  
26 positivity effect only occured when the moral valence are self-relevant but evidence to  
27 support such effect when the moral valence are other-relevant is weak. We further found  
28 that this effect exist even when the self-relevance or the moral valence were presented as a  
29 task-irrelevant information, though the effect size become much smaller. We also tested  
30 whether the positivity effect only exist in moral domain and found that this effect was not  
31 limited to moral domain. Exploratory analyses on task-questionnaire relationship found  
32 that moral self-image score (how closely one feel they are to the ideal moral image of  
33 themselves) is positively correlated to the  $d'$  of morally positive condition in singal  
34 detection and the drift rate using DDM, while the self-esteem is negatively correlated with  
35  $d'$  of neutral and morally negative conditions. These results suggest that the positive self  
36 prioritization in perceptual decision-making may reflect ...

37

*Keywords:* Perceptual decision-making, Self, positive bias, morality

38

Word count: X

39 Positivity bias in perceptual matching may reflect a spontaneous self-referential processing

40 **Introduction**

41 XXXX In perceptual matching, same is faster than different (Farell, 1985; Krueger,

42 1978). Automatic processing (Spruyt & Houwer, 2017)

43 Van Zandt, Colonius, and Proctor (2000): A comparison of two response time models

44 applied to perceptual matching

45 Yakushijin, ReikoJacobs, Robert A (2020), Are People Successful at Learning

46 Sequential Decisions on a Perceptual Matching Task?

47 Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model

48 for implicit effects in perceptual identification. Psychological Review, 108(1), 257–272.

49 <https://doi.org/10.1037/0033-295X.108.1.257>

50 We reported results from eleven experiments. In first set of experiments, we found

51 that shapes associated with morally positive person label were responded faster and more

52 accurately. In the second set of experiments, we explore the potential role of good self in

53 perceptual matching task and added one more independent variable, we found that the

54 effect was mainly on good self. In the third part we tested whether the morality will

55 automatically binds with person-relevance. Finally, we explore the correlation between

56 behavioral task and questionnaire scores.

57 **Disclosures**

58 We reported all the measurements, analyses, and results in all the experiments in the

59 current study. Participants whose overall accuracy lower than 60% were excluded from

60 analysis. Also, the accurate responses with less than 200ms reaction times were excluded

61 from the analysis.

62 All the experiments reported were not pre-registered. Most experiments (1a ~ 6b,  
63 except experiment 3b) reported in the current study were first finished between 2014 to  
64 2016 in Tsinghua University, Beijing, China. Participants in these experiments were  
65 recruited in the local community. To increase the sample size of experiments to 50 or more  
66 (Simmons, Nelson, & Simonsohn, 2013), we recruited additional participants in Wenzhou  
67 University, Wenzhou, China in 2017 for experiment 1a, 1b, 4a, and 4b. Experiment 3b was  
68 finished in Wenzhou University in 2017. To have a better estimation of the effect size, we  
69 included the data from two experiments (experiment 7a, 7b) that were reported in Hu,  
70 Lan, Macrae, and Sui (2020) (See Table S1 for overview of these experiments).

71 All participant received informed consent and compensated for their time. These  
72 experiments were approved by the ethic board in the Department of Tsinghua University.

### 73 General methods

#### 74 Design and Procedure

75 This series of experiments started to test the effect of instantly acquired true self  
76 (moral self) on perceptual decision-making. For this purpose, we used the social associative  
77 learning paradigm (or tagging paradigm)(Sui, He, & Humphreys, 2012), in which  
78 participants first learned the associations between geometric shapes and labels of person  
79 with different moral character (e.g., in first three studies, the triangle, square, and circle  
80 and good person, neutral person, and bad person, respectively). The associations of the  
81 shapes and label were counterbalanced across participants. After remembered the  
82 associations, participants finished a practice phase to familiar with the task, in which they  
83 viewed one of the shapes upon the fixation while one of the labels below the fixation and  
84 judged whether the shape and the label matched the association they learned. When  
85 participants reached 60% or higher accuracy at the end of the practicing session, they  
86 started the experimental task which was the same as in the practice phase.

87        The experiment 1a, 1b, 1c, 2, and 6a shared a 2 (matching: match vs. nonmatch) by  
88 3 (moral valence: good vs. neutral vs. bad) within-subject design. Experiment 1a was the  
89 first one of the whole series studies and 1b, 1c, and 2 were conducted to exclude the  
90 potential confounding factors. More specifically, experiment 1b used different Chinese  
91 words as label to test whether the effect only occurred with certain familiar words.  
92 Experiment 1c manipulated the moral valence indirectly: participants first learned to  
93 associate different moral behaviors with different neutral names, after remembered the  
94 association, they then performed the perceptual matching task by associating names with  
95 different shapes. Experiment 2 further tested whether the way we presented the stimuli  
96 influence the effect of valence, by sequentially presenting labels and shapes. Note that part  
97 of participants of experiment 2 were from experiment 1a because we originally planned a  
98 cross task comparison. Experiment 6a, which shared the same design as experiment 2, was  
99 an EEG experiment which aimed at exploring the neural correlates of the effect. But we  
100 will focus on the behavioral results of experiment 6a in the current manuscript.

101        For experiment 3a, 3b, 4a, 4b, 6b, 7a, and 7b, we included self-reference as another  
102 within-subject variable in the experimental design. For example, the experiment 3a directly  
103 extend the design of experiment 1a into a 2 (matchness: match vs. nonmatch) by 2  
104 (reference: self vs. other) by 3 (moral valence: good vs. neutral vs. bad) within-subject  
105 design. Thus in experiment 3a, there were six conditions (good-self, neutral-self, bad-self,  
106 good-other, neutral-other, and bad-other) and six shapes (triangle, square, circle, diamond,  
107 pentagon, and trapezoids). The experiment 6b was an EEG experiment extended from  
108 experiment 3a but presented the label and shape sequentially. Because of the relatively  
109 high working memory load (six label-shape pairs), experiment 6b were conducted in two  
110 days: the first day participants finished perceptual matching task as a practice, and the  
111 second day, they finished the task again while the EEG signals were recorded. Experiment  
112 3b was designed to separate the self-referential trials and other-referential trials. That is,  
113 participants finished two different blocks: in the self-referential blocks, they only responded

114 to good-self, neutral-self, and bad-self, with half match trials and half non-match trials; for  
115 the other-reference blocks, they only responded to good-other, neutral-other, and  
116 bad-other. Experiment 7a and 7b were designed to test the cross task robustness of the  
117 effect we observed in the aforementioned experiments (see, Hu et al., 2020). The matching  
118 task in these two experiments shared the same design with experiment 3a, but only with  
119 two moral valence, i.e., good vs. bad. We didn't include the neutral condition in  
120 experiment 7a and 7b because we found that the neutral and bad conditions constantly  
121 showed non-significant results in experiment 1 ~ 6.

122 Experiment 4a and 4b were design to test the automaticity of the binding between  
123 self/other and moral valence. In 4a, we used only two labels (self vs. other) and two shapes  
124 (circle, square). To manipulate the moral valence, we added the moral-related words within  
125 the shape and instructed participants to ignore the words in the shape during the task. In  
126 4b, we reversed the role of self-reference and valence in the task: participant learnt three  
127 labels (good-person, neutral-person, and bad-person) and three shapes (circle, square, and  
128 triangle), and the words related to identity, “self” or “other”, were presented in the shapes.  
129 As in 4a, participants were told to ignore the words inside the shape during the task.

130 Finally, experiment 5 was design to test the specificity of the moral valence. We  
131 extended experiment 1a with an additional independent variable: domains of the valence  
132 words. More specifically, besides the moral valence, we also added valence from other  
133 domains: appearance of person (beautiful, neutral, ugly), appearance of a scene (beautiful,  
134 neutral, ugly), and emotion (happy, neutral, and sad). Label-shape pairs from different  
135 domains were separated into different blocks.

136 E-prime 2.0 was used for presenting stimuli and collecting behavioral responses,  
137 except that experiment 7a and 7b used Matlab Psychtoolbox (Brainard, 1997; Pelli, 1997).  
138 For participants recruited in Tsinghua University, they finished the experiment individually  
139 in a dim-lighted chamber, stimuli were presented on 22-inch CRT monitors and their head

were fixed by a chin-rest brace. The distance between participants' eyes and the screen was about 60 cm. The visual angle of geometric shapes was about  $3.7^\circ \times 3.7^\circ$ , the fixation cross is of ( $0.8^\circ \times 0.8^\circ$  of visual angle) at the center of the screen. The words were of  $3.6^\circ \times 1.6^\circ$  visual angle. The distance between the center of the shape or the word and the fixation cross was  $3.5^\circ$  of visual angle. For participants recruited in Wenzhou University, they finished the experiment in a group consisted of 3 ~ 12 participants in a dim-lighted testing room. Participants were required to finished the whole experiment independently. Also, they were instructed to start the experiment at the same time, so that the distraction between participants were minimized. The stimuli were presented on 19-inch CRT monitor. The visual angles are could not be exactly controlled because participants's chin were not fixed.

In most of these experiments, participant were also asked to fill a battery of questionnaire after they finish the behavioral tasks. All the questionnaire data are open (see, dataset 4 in Liu et al., 2020). See Table S1 for a summary information about all the experiments.

## 155 Data analysis

156 **Analysis of individual study.** We used the `tidyverse` of r (see script  
157 `Load_save_data.r`) to exclude the practicing trials, invalid trials of each participants, and  
158 invalid participants, if there were any, in the raw data. Results of each experiment were  
159 then analyzed in three different approaches.

### 160 *Classic NHST.*

161 First, as in Sui et al. (2012), we analyzed the accuracy and reaction times using  
162 classic repeated measures ANOVA in the Null Hypothesis Significance Test (NHST)  
163 framework. Repeated measures ANOVAs is essentially a two-step mixed model. In the first  
164 step, we estimate the parameter on individual level, and in the second step, we used

165 repeated ANOVA to test the Null hypothesis. More specifically, for the accuracy, we used a  
 166 signal detection approach, in which individual' sensitivity  $d'$  was estimated first. To  
 167 estimate the sensitivity, we treated the match condition as the signal while the nonmatch  
 168 conditions as noise. Trials without response were coded either as “miss” (match trials) or  
 169 “false alarm” (nonmatch trials). Given that the match and nonmatch trials are presented  
 170 in the same way and had same number of trials across all studies, we assume that  
 171 participants' inner distribution of these two types of trials had equal variance but may had  
 172 different means. That is, we used the equal variance Gaussian SDT model (EVSDT) here  
 173 (Rouder & Lu, 2005). The  $d'$  was then estimated as the difference of the standardized hit  
 174 and false alarm rats (Stanislaw & Todorov, 1999):

$$d' = zHR - zFAR = \Phi^{-1}(HR) - \Phi^{-1}(FAR)$$

175 where the  $HR$  means hit rate and the  $FAR$  mean false alarm rate.  $zHR$  and  $zFAR$  are  
 176 the standardized hit rate and false alarm rates, respectively. These two  $z$ -scores were  
 177 converted from proportion (i.e., hit rate or false alarm rate) by inverse cumulative normal  
 178 density function,  $\Phi^{-1}$  ( $\Phi$  is the cumulative normal density function, and is used convert  $z$   
 179 score into probabilities). Another parameter of signal detection theory, response criterion  $c$ ,  
 180 is defined by the negative standardized false alarm rate (DeCarlo, 1998):  $-zFAR$ .

181 For the reaction times (RTs), only RTs of accurate trials were analyzed. We first  
 182 calculate the mean RTs of each participant and then subject the mean RTs of each  
 183 participant to repeated measures ANOVA. Note that we set the alpha as .05. The repeated  
 184 measure ANOVA was done by `afex` package (<https://github.com/singmann/afex>).

185 To control the false positive rate when conducting the post-hoc comparisons, we used  
 186 Bonferroni correction.

187 ***Bayesian hierarchical generalized linear model (GLM).***

188 The classic NHST approach may ignore the uncertainty in estimate of the parameters  
 189 for SDT (Rouder & Lu, 2005), and using mean RT assumes normal distribution of RT

190 data, which is always not true because RTs distribution is skewed (Rousselet & Wilcox,  
 191 2019). To better estimate the uncertainty and use a more appropriate model, we also tried  
 192 Bayesian hierarchical generalized linear model to analyze each experiment's accuracy and  
 193 RTs data. We used BRMs (Bürkner, 2017) to build the model, which used Stan (Carpenter  
 194 et al., 2017) to estimate the posterior.

195 In the GLM model, we assume that the accuracy of each trial is Bernoulli distributed  
 196 (binomial with 1 trial), with probability  $p_i$  that  $y_i = 1$ .

$$y_i \sim \text{Bernoulli}(p_i)$$

197 In the perceptual matching task, the probability  $p_i$  can then be modeled as a function of  
 198 the trial type:

$$\Phi(p_i) = \beta_0 + \beta_1 \text{IsMatch}_i * \text{Valence}_i$$

199 The outcomes  $y_i$  are 0 if the participant responded “nonmatch” on trial  $i$ , 1 if they  
 200 responded “match”. The probability of the “match” response for trial  $i$  for a participant is  
 201  $p_i$ . We then write the generalized linear model on the probits (z-scores;  $\Phi$ , “Phi”) of  $ps$ .  $\Phi$   
 202 is the cumulative normal density function and maps  $z$  scores to probabilities. Given this  
 203 parameterization, the intercept of the model ( $\beta_0$ ) is the standardized false alarm rate  
 204 (probability of saying 1 when predictor is 0), which we take as our criterion  $c$ . The slope of  
 205 the model ( $\beta_1$ ) is the increase of saying 1 when predictor is 1, in  $z$ -scores, which is another  
 206 expression of  $d'$ . Therefore,  $c = -z\text{HR} = -\beta_0$ , and  $d' = \beta_1$ .

207 In each experiment, we had multiple participants, then we need also consider the  
 208 variations between subjects, i.e., a hierarchical mode in which individual's parameter and  
 209 the population level parameter are estimated simultaneously. We assume that the  
 210 outcome of each trial is Bernoulli distributed (binomial with 1 trial), with probability  $p_{ij}$   
 211 that  $y_{ij} = 1$ .

$$y_{ij} \sim Bernoulli(p_{ij})$$

<sup>212</sup> Similarly, the generalized linear model was extended to two levels:

$$\Phi(p_{ij}) = \beta_{0j} + \beta_{1j} IsMatch_{ij} * Valence_{ij}$$

- <sup>213</sup> The outcomes  $y_{ij}$  are 0 if participant  $j$  responded “nonmatch” on trial  $i$ , 1 if they  
<sup>214</sup> responded “match”. The probability of the “match” response for trial  $i$  for subject  $j$  is  $p_{ij}$ .  
<sup>215</sup> We again can write the generalized linear model on the probits (z-scores;  $\Phi$ , “Phi”) of  $ps$ .

<sup>216</sup> The subjective-specific intercepts ( $\beta_0 = -zFAR$ ) and slopes ( $\beta_1 = d'$ ) are described  
<sup>217</sup> by multivariate normal with means and a covariance matrix for the parameters.

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}, \Sigma\right)$$

<sup>218</sup> For the reaction time, we used the log normal distribution

<sup>219</sup> ([https://lindeloev.github.io/shiny-rt/#34\\_\(shifted\)\\_log-normal](https://lindeloev.github.io/shiny-rt/#34_(shifted)_log-normal)). This distribution has  
<sup>220</sup> two parameters:  $\mu$ ,  $\sigma$ .  $\mu$  is the mean of the logNormal distribution, and  $\sigma$  is the disperse of  
<sup>221</sup> the distribution. The log normal distribution can be extended to shifted log normal  
<sup>222</sup> distribution, with one more parameter: shift, which is the earliest possible response.

$$y_i = \beta_0 + \beta_1 * IsMatch_i * Valence_i$$

<sup>223</sup> Shifted log-normal distribution:

$$\log(y_{ij}) \sim N(\mu_j, \sigma_j)$$

<sup>224</sup>  $y_{ij}$  is the RT of the  $i$ th trial of the  $j$ th participants.

$$\mu_j \sim N(\mu, \sigma)$$

$$\sigma_j \sim Cauchy()$$

225        ***Hierarchical drift diffusion model (HDDM).***

226        To further explore the psychological mechanism under perceptual decision-making, we  
 227        used HDDM (Wiecki, Sofer, & Frank, 2013) to model our RTs and accuracy data. We used  
 228        the prior implemented in HDDM, that is, informative priors that constrains parameter  
 229        estimates to be in the range of plausible values based on past literature (Matzke &  
 230        Wagenmakers, 2009). As reported in Hu et al. (2020), we used the response code approach,  
 231        match response were coded as 1 and nonmatch responses were coded as 0. To fully explore  
 232        all parameters, we allow all four parameters of DDM free to vary. We then extracted the  
 233        estimation of all the four parameters for each participants for the correlation analyses.  
 234        However, because the starting point is only related to response (match vs. non-match) but  
 235        not the valence of the stimuli, we didn't included it in correlation analysis.

236        **Synthesized results.** We also reported the synthesized results from the  
 237        experiments, because many of them shared the similar experimental design. We reported  
 238        the results in five parts: valence effect, explicit interaction between valence and  
 239        self-relevance, implicit interaction between valence and self-relevance, specificity of valence  
 240        effect, and behavior-questionnaire correlation.

241        For the first two parts, we reported the synthesized results from Frequentist's  
 242        approach(mini-meta-analysis, Goh, Hall, & Rosenthal, 2016). The mini meta-analyses were  
 243        carried out by using `metafor` package (Viechtbauer, 2010). We first calculated the mean of  
 244         $d'$  and RT of each condition for each participant, then calculate the effect size (Cohen's  $d$ )  
 245        and variance of the effect size for all contrast we interested: Good v. Bad, Good v.  
 246        Neutral, and Bad v. Neutral for the effect of valence, and self vs. other for the effect of  
 247        self-relevance. Cohen's  $d$  and its variance were estimated using the following formula  
 248        (Cooper, Hedges, & Valentine, 2009):

$$d = \frac{(M_1 - M_2)}{\sqrt{(sd_1^2 + sd_2^2) - 2rsd_1sd_2}} \sqrt{2(1-r)}$$

$$var.d = 2(1 - r)(\frac{1}{n} + \frac{d^2}{2n})$$

<sup>249</sup>  $M_1$  is the mean of the first condition,  $sd_1$  is the standard deviation of the first  
<sup>250</sup> condition, while  $M_2$  is the mean of the second condition,  $sd_2$  is the standard deviation of  
<sup>251</sup> the second condition.  $r$  is the correlation coefficient between data from first and second  
<sup>252</sup> condition.  $n$  is the number of data point (in our case the number of participants included  
<sup>253</sup> in our research).

<sup>254</sup> The effect size from each experiment were then synthesized by random effect model  
<sup>255</sup> using `metafor` (Viechtbauer, 2010). Note that to avoid the cases that some participants  
<sup>256</sup> participated more than one experiments, we inspected the all available information of  
<sup>257</sup> participants and only included participants' results from their first participation. As  
<sup>258</sup> mentioned above, 24 participants were intentionally recruited to participate both exp 1a  
<sup>259</sup> and exp 2, we only included their results from experiment 1a in the meta-analysis.

<sup>260</sup> We also estimated the synthesized effect size using Bayesian hierarchical model,  
<sup>261</sup> which extended the two-level hierarchical model in each experiment into three-level model,  
<sup>262</sup> which experiment as an additional level. For SDT, we can use a nested hierarchical model  
<sup>263</sup> to model all the experiment with similar design:

$$y_{ijk} \sim Bernoulli(p_{ijk})$$

<sup>264</sup> where

$$\Phi(p_{ijk}) = \beta_{0jk} + \beta_{1jk} IsMatch_{ijk}$$

<sup>265</sup> The outcomes  $y_{ijk}$  are 0 if participant  $j$  in experiment k responded “nonmatch” on trial  $i$ ,  
<sup>266</sup> 1 if they responded “match”.

$$\begin{bmatrix} \beta_{0jk} \\ \beta_{1jk} \end{bmatrix} \sim N\left(\begin{bmatrix} \theta_{0k} \\ \theta_{1k} \end{bmatrix}, \sum\right)$$

<sup>267</sup> and the experiment level parameter  $mu_{0k}$  and  $mu_{1k}$  is from a higher order  
<sup>268</sup> distribution:

$$\begin{bmatrix} \theta_{0k} \\ \theta_{1k} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \Sigma\right)$$

<sup>269</sup> in which  $\mu_0$  and  $\mu_1$  means the population level parameter.

<sup>270</sup> This model can be easily expand to three-level model in which participants and  
<sup>271</sup> experiments are two group level variable and participants were nested in the experiments.

$$\log(y_{ijk}) \sim N(\mu_{jk}, \sigma_{jk})$$

<sup>272</sup>  $y_{ijk}$  is the RT of the  $i$ th trial of the  $j$ th participants in the  $k$ th experiment.

$$\mu_{jk} \sim N(\mu_k, \sigma_k)$$

$$\sigma_{jk} \sim Cauchy()$$

$$\theta_{jk} \sim Cauchy()$$

$$\mu_k \sim N(\mu, \sigma)$$

<sup>273</sup> Using the Bayesian hierarchical model, we can directly estimate the over-all effect of  
<sup>274</sup> valence on  $d'$  across all experiments with similar experimental design, instead of using a  
<sup>275</sup> two-step approach where we first estimate the  $d'$  for each participant and then use a  
<sup>276</sup> random effect model meta-analysis (Goh et al., 2016).

277        ***Valence effect.***

278        We synthesized effect size of  $d'$  and RT from experiment 1a, 1b, 1c, 2, 5 and 6a for  
279        the valence effect. We reported the synthesized the effect across all experiments that tested  
280        the valence effect, using the mini meta-analysis approach (Goh et al., 2016).

281        ***Explicit interaction between Valence and self-relevance.***

282        The results from experiment 3a, 3b, 6b, 7a, and 7b. These experiments explicitly  
283        included both moral valence and self-reference.

284        ***Implicit interaction between valence and self-relevance.***

285        In the third part, we focused on experiment 4a and 4b, which were designed to  
286        examine the implicit effect of the interaction between moral valence and self-referential  
287        processing. We are interested in one particular question: will self-referential and morally  
288        positive valence had a mutual facilitation effect. That is, when moral valence (experiment  
289        4a) or self-referential (experiment 4a) was presented as task-irrelevant stimuli, whether  
290        they would facilitate self-referential or valence effect on perceptual decision-making. For  
291        experiment 4a, we reported the comparisons between different valence conditions under the  
292        self-referential task and other-referential task. For experiment 4b, we first calculated the  
293        effect of valence for both self- and other-referential conditions and then compared the effect  
294        size of these three contrast from self-referential condition and from other-referential  
295        condition. Note that the results were also analyzed in a standard repeated measure  
296        ANOVA (see supplementary materials).

297        ***Specificity of the valence effect.***

298        In this part, we reported the data from experiment 5, which included positive,  
299        neutral, and negative valence from four different domains: morality, aesthetic of person,  
300        aesthetic of scene, and emotion. This experiment was design to test whether the positive  
301        bias is specific to morality.

302        ***Behavior-Questionnaire correlation.***

303        Finally, we explored correlation between results from behavioral results and  
 304        self-reported measures.

305        For the questionnaire part, we are most interested in the self-rated distance between  
 306        different person and self-evaluation related questionnaires: self-esteem, moral-self identity,  
 307        and moral self-image. Other questionnaires (e.g., personality) were not planned to  
 308        correlated with behavioral data were not included. Note that all data were reported in (Liu  
 309        et al., 2020).

310        For the behavioral task part, we derived different indices. First, we used the mean of  
 311        the RT and  $d'$  from each participants of each condition. Second, we used three parameters  
 312        from drift diffusion model: drift rate ( $v$ ), boundary separation ( $a$ ), and non  
 313        decision-making time ( $t$ ). Third, we calculated the differences between different conditions  
 314        (valence effect: good-self vs. bad-self, good-self vs. neutral-self, bad-self vs. neutral-self;  
 315        good-other vs. bad-other, good-other vs. neutral-other, bad-other vs. neutral-other;  
 316        Self-reference effect: good-self vs. good-other, neutral-self vs. neutral-other, bad-self  
 317        vs. bad-other), as indexed by Cohen's  $d$  and standard error (SE) of Cohen's  $d$ .

$$Cohen's d_z = \frac{(M_1 - M_2)}{\sqrt{(SD_1^2 + SD_2^2)/2}}$$

318        Given that the task difficulty were different across experiments, we z-transformed all these  
 319        indices so that they become unit-free.

320        We used the mean of parameter posterior distribution as the estimate of each  
 321        parameter for each participants in the correlation analysis.

322        We used Pearson correlation to quantify the correlation. For those correlation that is  
 323        significant ( $p < 0.05$ ), we further tested the robustness of the correlation using bootstrap  
 324        by BootES package (Kirby & Gerlanc, 2013). To avoid false positive, we further determined  
 325        the threshold for significant by permutation. More specifically, for each pairs that initially

326 with  $p < .05$ , we randomly shuffle the participants data of each score and calculated the  
327 correlation between the shuffled vectors. After repeating this procedure for 5000 times, we  
328 choose arrange these 5000 correlation coefficients and use the 95% percentile number as our  
329 threshold.

330 **Part 1: Moral valence effect**

331 In this part, we report five experiments that aimed at testing whether the instantly  
332 acquired association between shapes and good person would be prioritized in perceptual  
333 decision-making.

334 **Experiment 1a**

335 **Methods.**

336 ***Participants.***

337 57 college students (38 female, age =  $20.75 \pm 2.54$  years) participated. 39 of them  
338 were recruited from Tsinghua University community in 2014; 18 were recruited from  
339 Wenzhou University in 2017. All participants were right-handed except one, and all had  
340 normal or corrected-to-normal vision. Informed consent was obtained from all participants  
341 prior to the experiment according to procedures approved by the local ethics committees. 6  
342 participant's data were excluded from analysis because nearly random level of accuracy,  
343 leaving 51 participants (34 female, age =  $20.72 \pm 2.44$  years).

344 ***Stimuli and Tasks.***

345 Three geometric shapes were used in this experiment: triangle, square, and circle.  
346 These shapes were paired with three labels (bad person, good person or neutral person).  
347 The pairs were counterbalanced across participants.

348      ***Procedure.***

349      This experiment had two phases. First, there was a brief learning stage. Participants  
350     were asked to learn the relationship between geometric shapes (triangle, square, and circle)  
351     and different person (bad person, a good person, or a neutral person). For example, a  
352     participant was told, “bad person is a circle; good person is a triangle; and a neutral person  
353     is represented by a square.” After participant remember the associations (usually in a few  
354     minutes), participants started a practicing phase of matching task which has the exact task  
355     as in the experimental task. In the experimental task, participants judged whether  
356     shape–label pairs, which were subsequently presented, were correct. Each trial started with  
357     the presentation of a central fixation cross for 500 ms. Subsequently, a pairing of a shape  
358     and label (good person, bad person, and neutral person) was presented for 100 ms. The  
359     pair presented could confirm to the verbal instruction for each pairing given in the training  
360     stage, or it could be a recombination of a shape with a different label, with the shape–label  
361     pairings being generated at random. The next frame showed a blank for 1100ms.

362      Participants were expected to judge whether the shape was correctly assigned to the person  
363     by pressing one of the two response buttons as quickly and accurately as possible within  
364     this timeframe (to encourage immediate responding). Feedback (correct or incorrect) was  
365     given on the screen for 500 ms at the end of each trial, if no response detected, “too slow”  
366     was presented to remind participants to accelerate. Participants were informed of their  
367     overall accuracy at the end of each block. The practice phase finished and the experimental  
368     task began after the overall performance of accuracy during practice phase achieved 60%.

369      For participants from the Tsinghua community, they completed 6 experimental blocks of 60  
370     trials. Thus, there were 60 trials in each condition (bad-person match, bad-person  
371     nonmatch, good-person match, good-person nonmatch, neutral-person match, and  
372     neutral-person nonmatch). For the participants from Wenzhou University, they finished 6  
373     blocks of 120 trials, therefore, 120 trials for each condition.

374      ***Data analysis.***

375 As described in general methods section, this experiment used three approaches to  
376 analyze the behavioral data: Classical NHST, Bayesian Hierarchical Generalized Linear  
377 Model, and Hierarchical drift diffusion model.

378 **Results.**

379 ***Classic NHST.***

380 *d prime.*

381 Figure 1 shows *d* prime and reaction times during the perceptual matching task. We  
382 conducted a single factor (valence: good, neutral, bad) repeated measure ANOVA.

383 We found the effect of Valence ( $F(1.96, 97.84) = 6.19$ ,  $MSE = 0.27$ ,  $p = .003$ ,  
384  $\hat{\eta}_G^2 = .020$ ). The post-hoc comparison with multiple comparison correction revealed that  
385 the shapes associated with Good-person (2.11, SE = 0.14) has greater *d* prime than shapes  
386 associated with Bad-person (1.75, SE = 0.14),  $t(50) = 3.304$ ,  $p = 0.0049$ . The Good-person  
387 condition was also greater than the Neutral-person condition (1.95, SE = 0.16), but didn't  
388 reach statistical significant,  $t(50) = 1.54$ ,  $p = 0.28$ . Neither the Neutral-person condition is  
389 significantly greater than the Bad-person condition,  $t(50) = 2.109$ ,  $p = .098$ .

390 *Reaction times.*

391 We conducted 2 (Matchness: match v. nonmatch) by 3 (Valence: good, neutral, bad)  
392 repeated measure ANOVA. We found the main effect of Matchness ( $F(1, 50) = 232.39$ ,  
393  $MSE = 948.92$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .104$ ), main effect of valence ( $F(1.87, 93.31) = 9.62$ ,  
394  $MSE = 1,673.86$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .016$ ), and interaction between Matchness and Valence  
395 ( $F(1.73, 86.65) = 8.52$ ,  $MSE = 1,441.75$ ,  $p = .001$ ,  $\hat{\eta}_G^2 = .011$ ).

396 We then carried out two separate ANOVA for Match and Mismatched trials. For  
397 matched trials, we found the effect of valence . We further examined the effect of valence  
398 for both self and other for matched trials. We found that shapes associated with Good  
399 Person (684 ms, SE = 11.5) responded faster than Neutral (709 ms, SE = 11.5),  $t(50) =$

400 -2.265,  $p = 0.0702$ ) and Bad Person (728 ms, SE = 11.7),  $t(50) = -4.41$ ,  $p = 0.0002$ ), and  
 401 the Neutral condition was faster than the Bad condition,  $t(50) = -2.495$ ,  $p = 0.0415$ ). For  
 402 non-matched trials, there was no significant effect of Valence ().

403 ***Bayesian hierarchical GLM.***

404 *d prime.*

405 We fitted a Bayesian hierarchical GLM for signal detection theory approach. The  
 406 results showed that when the shapes were tagged with labels with different moral valence,  
 407 the sensitivity ( $d'$ ) and criteria ( $c$ ) were both influence. For the  $d'$ , we found that the  
 408 shapes tagged with morally good person (2.46, 95% CI[2.21 2.72]) is greater than shapes  
 409 tagged with moral bad (2.07, 95% CI[1.83 2.32]),  $P_{PosteriorComparison} = 1$ . Shape tagged  
 410 with morally good person is also greater than shapes tagged with neutral person (2.23,  
 411 95% CI[1.95 2.49]),  $P_{PosteriorComparison} = 0.97$ . Also, the shapes tagged with neutral  
 412 person is greater than shapes tagged with morally bad person,  $P_{PosteriorComparison} = 0.92$ .

413 Interesting, we also found the criteria for three conditions also differ, the shapes  
 414 tagged with good person has the highest criteria (-1.01, [-1.14 -0.88]), followed by shapes  
 415 tagged with neutral person(-1.06, [-1.21 -0.92]), and then the shapes tagged with bad  
 416 person(-1.11, [-1.25 -0.97]). However, pair-wise comparison showed that only showed strong  
 417 evidence for the difference between good and bad conditions.

418 *Reaction times.*

419 We fitted a Bayesian hierarchical GLM for RTs, with a log-normal distribution as the  
 420 link function. We used the posterior distribution of the regression coefficient to make  
 421 statistical inferences. As in previous studies, the matched conditions are much faster than  
 422 the mismatched trials ( $P_{PosteriorComparison} = 1$ ). We focused on matched trials only, and  
 423 compared different conditions: Good is faster than the neutral,  $P_{PosteriorComparison} = .99$ ,  
 424 it was also faster than the Bad condition,  $P_{PosteriorComparison} = 1$ . And the neutral  
 425 condition is faster than the bad condition,  $P_{PosteriorComparison} = .99$ . However, the

426 mismatched trials are largely overlapped. See Figure 2.

427 **HDDM.**

428 We fitted our data with HDDM, using the response-coding (See also, Hu et al., 2020).

429 We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary separation ( $a$ )  
430 for each condition. We found that the shapes tagged with good person has higher drift rate  
431 and higher boundary separation than shapes tagged with both neutral and bad person.

432 Also, the shapes tagged with neutral person has a higher drift rate than shapes tagged  
433 with bad person, but not for the boundary separation. Finally, we found that shapes  
434 tagged with bad person had longer non-decision time (see Figure 3).

435 **Experiment 1b**

436 In this study, we aimed at excluding the potential confounding factor of the

437 familiarity of words we used in experiment 1a, by matching the familiarity of the words.

438 **Method.**

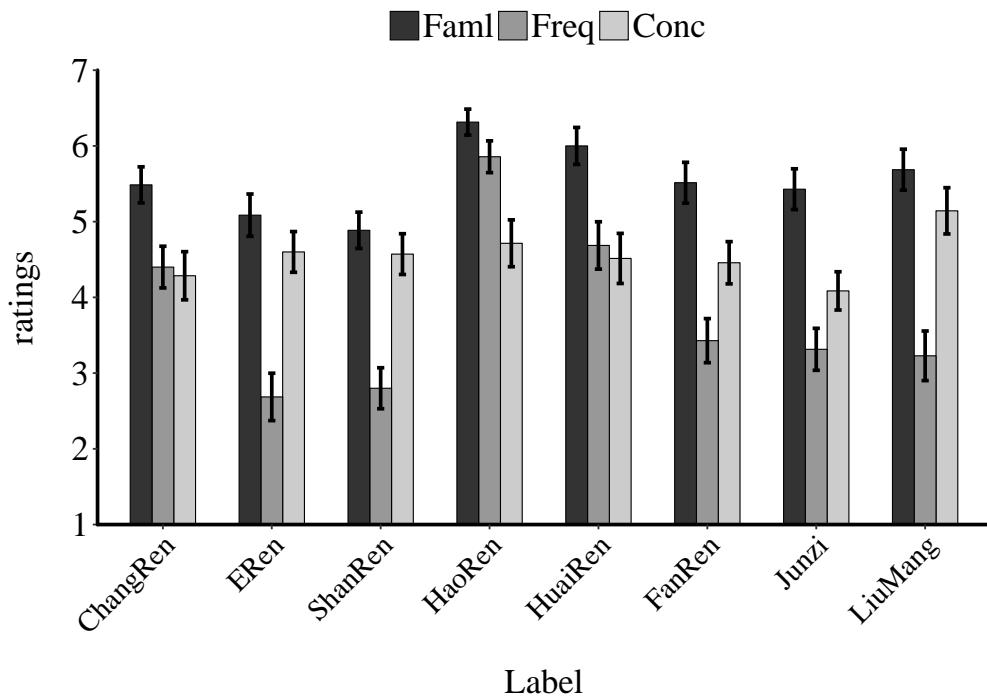
439 **Participants.**

440 72 college students (49 female, age =  $20.17 \pm 2.08$  years) participated. 39 of them  
441 were recruited from Tsinghua University community in 2014; 33 were recruited from  
442 Wenzhou University in 2017. All participants were right-handed except one, and all had  
443 normal or corrected-to-normal vision. Informed consent was obtained from all participants  
444 prior to the experiment according to procedures approved by the local ethics committees.  
445 20 participant's data were excluded from analysis because nearly random level of accuracy,  
446 leaving 52 participants (36 female, age =  $20.25 \pm 2.31$  years).

447 **Stimuli and Tasks.** Three geometric shapes (triangle, square, and circle, with  $3.7^\circ$   
448  $\times 3.7^\circ$  of visual angle) were presented above a white fixation cross subtending  $0.8^\circ \times 0.8^\circ$   
449 of visual angle at the center of the screen. The three shapes were randomly assigned to  
450 three labels with different moral valence: a morally bad person (" ", ERen), a morally

451 good person (“ ”, ShanRen) or a morally neutral person (“ ”, ChangRen). The order of  
 452 the associations between shapes and labels was counterbalanced across participants. Three  
 453 labels used in this experiment is selected based on the rating results from an independent  
 454 survey, in which participants rated the familiarity, frequency, and concreteness of eight  
 455 different words online. Of the eight words, three of them are morally positive (HaoRen,  
 456 ShanRen, Junzi), two of them are morally neutral (ChangRen, FanRen), and three of them  
 457 are morally negative (HuaiRen, ERen, LiuMang). An independent sample consist of 35  
 458 participants (22 females, age  $20.6 \pm 3.11$ ) were recruited to rate these words. Based on the  
 459 ratings (see supplementary materials Figure S1), we selected ShanRen, ChangRen, and  
 460 ERen to represent morally positive, neutral, and negative person.

#### Ratings for each label



461

#### *Procedure.*

462 For participants from both Tsinghua community and Wenzhou community, the  
 463 procedure in the current study was exactly same as in experiment 1a.

464

465 **Data Analysis.** Data was analyzed as in experiment 1a.

466        **Results.**

467        **NHST.**

468        Figure 4 shows  $d$  prime and reaction times of experiment 1b.

469         $d$  prime.

470        Repeated measures ANOVA revealed main effect of valence,  $F(1.83, 93.20) = 14.98$ ,

471         $MSE = 0.18$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .053$ . Paired t test showed that the Good-Person condition

472         $(1.87 \pm 0.102)$  was with greater  $d$  prime than Neutral condition  $(1.44 \pm 0.101$ ,  $t(51) =$

473         $5.945$ ,  $p < 0.001$ ). We also found that the Bad-Person condition  $(1.67 \pm 0.11)$  has also

474        greater  $d$  prime than neutral condition ,  $t(51) = 3.132$ ,  $p = 0.008$ ). There Good-person

475        condition was also slightly greater than the bad condition,  $t(51) = 2.265$ ,  $p = 0.0701$ .

476        *Reaction times.*

477        We found interaction between Matchness and Valence ( $F(1.95, 99.31) = 19.71$ ,

478         $MSE = 960.92$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .031$ ) and then analyzed the matched trials and

479        mismatched trials separately, as in experiment 1a. For matched trials, we found the effect

480        of valence  $F(1.94, 99.10) = 33.97$ ,  $MSE = 1,343.19$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .115$ . Post-hoc  $t$ -tests

481        revealed that shapes associated with Good Person  $(684 \pm 8.77)$  were responded faster than

482        Neutral-Person  $(740 \pm 9.84)$ ,  $(t(51) = -8.167$ ,  $p < 0.001)$  and Bad Person  $(728 \pm 9.15)$ ,

483         $t(51) = -5.724$ ,  $p < 0.0001$ ). While there was no significant differences between Neutral and

484        Bad-Person condition  $(t(51) = 1.686$ ,  $p = 0.221$ ). For non-matched trials, there was no

485        significant effect of Valence ( $F(1.90, 97.13) = 1.80$ ,  $MSE = 430.15$ ,  $p = .173$ ,  $\hat{\eta}_G^2 = .003$ ).

486        **BGLM.**

487        *Signal detection theory analysis of accuracy.*

488        We fitted a Bayesian hierarchical GLM for SDT. The results showed that when the

489        shapes were tagged with labels with different moral valence, the sensitivity ( $d'$ ) and criteria

490        ( $c$ ) were both influence. For the  $d'$ , we found that the shapes tagged with morally good

491 person (2.46, 95% CI[2.21 2.72]) is greater than shapes tagged with moral bad (2.07, 95%  
 492 CI[1.83 2.32]),  $P_{PosteriorComparison} = 1$ . Shape tagged with morally good person is also  
 493 greater than shapes tagged with neutral person (2.23, 95% CI[1.95 2.49]),  
 494  $P_{PosteriorComparison} = 0.97$ . Also, the shapes tagged with neutral person is greater than  
 495 shapes tagged with morally bad person,  $P_{PosteriorComparison} = 0.92$ .

496 Interesting, we also found the criteria for three conditions also differ, the shapes  
 497 tagged with good person has the highest criteria (-1.01, [-1.14 -0.88]), followed by shapes  
 498 tagged with neutral person(1.06, [-1.21 -0.92]), and then the shapes tagged with bad  
 499 person(-1.11, [-1.25 -0.97]). However, pair-wise comparison showed that only showed strong  
 500 evidence for the difference between good and bad conditions.

501 *Reaction time.*

502 We fitted a Bayesian hierarchical GLM for RTs, with a log-normal distribution as the  
 503 link function. We used the posterior distribution of the regression coefficient to make  
 504 statistical inferences. As in previous studies, the matched conditions are much faster than  
 505 the mismatched trials ( $P_{PosteriorComparison} = 1$ ). We focused on matched trials only, and  
 506 compared different conditions: Good is faster than the neutral,  $P_{PosteriorComparison} = .99$ ,  
 507 it was also faster than the Bad condition,  $P_{PosteriorComparison} = 1$ . And the neutral  
 508 condition is faster than the bad condition,  $P_{PosteriorComparison} = .99$ . However, the  
 509 mismatched trials are largely overlapped. See Figure 5.

510 **HDDM.**

511 We found that the shapes tagged with good person has higher drift rate and higher  
 512 boundary separation than shapes tagged with both neutral and bad person. Also, the  
 513 shapes tagged with neutral person has a higher drift rate than shapes tagged with bad  
 514 person, but not for the boundary separation. Finally, we found that shapes tagged with  
 515 bad person had longer non-decision time (see figure 6).

516        **Discussion.** These results confirmed the facilitation effect of positive moral valence  
517        on the perceptual matching task. This pattern of results mimic prior results demonstrating  
518        self-bias effect on perceptual matching (Sui et al., 2012) and in line with previous studies  
519        that indirect learning of other's moral reputation do have influence on our subsequent  
520        behavior (Fouragnan et al., 2013).

521        **Experiment 1c**

522        In this study, we further control the valence of words using in our experiment.

523        Instead of using label with moral valence, we used valence-neutral names in China.  
524        Participant first learn behaviors of the different person, then, they associate the names and  
525        shapes. And then they perform a name-shape matching task.

526        **Method.**

527        ***Participants.***

528        23 college students (15 female, age =  $22.61 \pm 2.62$  years) participated. All of them  
529        were recruited from Tsinghua University community in 2014. Informed consent was  
530        obtained from all participants prior to the experiment according to procedures approved by  
531        the local ethics committees. No participant was excluded because they overall accuracy  
532        were above 0.6.

533        ***Stimuli and Tasks.***

534        Three geometric shapes (triangle, square, and circle, with  $3.7^\circ \times 3.7^\circ$  of visual angle)  
535        were presented above a white fixation cross subtending  $0.8^\circ \times 0.8^\circ$  of visual angle at the  
536        center of the screen. The three most common names were chosen, which are neutral in  
537        moral valence before the manipulation. Three names (Zhang, Wang, Li) were first paired  
538        with three paragraphs of behavioral description. Each description includes one sentence of  
539        biographic information and four sentences that describing the moral behavioral under that  
540        name. To assess the that these three descriptions represented good, neutral, and bad

541 valence, we collected the ratings of three person on six dimensions: morality, likability,  
542 trustworthiness, dominance, competence, and aggressiveness, from an independent sample  
543 ( $n = 34$ , 18 female, age =  $19.6 \pm 2.05$ ). The rating results showed that the person with  
544 morally good behavioral description has higher score on morality ( $M = 3.59$ ,  $SD = 0.66$ )  
545 than neutral ( $M = 0.88$ ,  $SD = 1.1$ ),  $t(33) = 12.94$ ,  $p < .001$ , and bad conditions ( $M = -3.4$ ,  
546  $SD = 1.1$ ),  $t(33) = 30.78$ ,  $p < .001$ . Neutral condition was also significant higher than bad  
547 conditions  $t(33) = 13.9$ ,  $p < .001$  (See supplementary materials).

548 **Procedure.**

549 After arriving the lab, participants were informed to complete two experimental  
550 tasks, first a social memory task to remember three person and their behaviors, after tested  
551 for their memory, they will finish a perceptual matching task. In the social memory task,  
552 the descriptions of three person were presented without time limitation. Participant  
553 self-paced to memorized the behaviors of each person. After they memorizing, a  
554 recognition task was used to test their memory effect. Each participant was required to  
555 have over 95% accuracy before preceding to matching task. The perceptual learning task  
556 was followed, three names were randomly paired with geometric shapes. Participants were  
557 required to learn the association and perform a practicing task before they start the formal  
558 experimental blocks. They kept practicing until they reached 70% accuracy. Then, they  
559 would start the perceptual matching task as in experiment 1a. They finished 6 blocks of  
560 perceptual matching trials, each have 120 trials.

561 **Data Analysis.** Data was analyzed as in experiment 1a.

562 **Results.** Figure 7 shows  $d$  prime and reaction times of experiment 1c. We  
563 conducted same analysis as in Experiment 1a. Our analysis didn't show effect of valence  
564 on  $d$  prime,  $F(1.93, 42.56) = 0.23$ ,  $MSE = 0.41$ ,  $p = .791$ ,  $\hat{\eta}_G^2 = .005$ . Neither the effect of  
565 valence on RT ( $F(1.63, 35.81) = 0.22$ ,  $MSE = 2,212.71$ ,  $p = .761$ ,  $\hat{\eta}_G^2 = .001$ ) or  
566 interaction between valence and matchness on RT ( $F(1.79, 39.43) = 1.20$ ,

<sup>567</sup>  $MSE = 1,973.91$ ,  $p = .308$ ,  $\hat{\eta}_G^2 = .005$ ).

<sup>568</sup> ***Signal detection theory analysis of accuracy.***

<sup>569</sup> We fitted a Bayesian hierarchical GLM for SDT. The results showed that when the  
<sup>570</sup> shapes were tagged with labels with different moral valence, the sensitivity ( $d'$ ) and criteria  
<sup>571</sup> ( $c$ ) were both influenced. For the  $d'$ , we found that the shapes tagged with morally good  
<sup>572</sup> person (2.30, 95% CI[1.93 2.70]) is greater than shapes tagged with moral bad (2.11, 95%  
<sup>573</sup> CI[1.83 2.42]),  $P_{PosteriorComparison} = 0.8$ . Shape tagged with morally good person is also  
<sup>574</sup> greater than shapes tagged with neutral person (2.16, 95% CI[1.88 2.45]),  
<sup>575</sup>  $P_{PosteriorComparison} = 0.75$ .

<sup>576</sup> Interesting, we also found the criteria for three conditions also differ, the shapes  
<sup>577</sup> tagged with good person has the highest criteria (-0.97, [-1.12 -0.82]), followed by shapes  
<sup>578</sup> tagged with neutral person(-0.96, [-1.09 -0.83]), and then the shapes tagged with bad  
<sup>579</sup> person(-1.03, [-1.22 -0.84]). However, pair-wise comparison showed that only showed strong  
<sup>580</sup> evidence for the difference between good and bad conditions.

<sup>581</sup> ***Reaction time.***

<sup>582</sup> We fitted a Bayesian hierarchical GLM for RTs, with a log-normal distribution as the  
<sup>583</sup> link function. We used the posterior distribution of the regression coefficient to make  
<sup>584</sup> statistical inferences. As in previous studies, the matched conditions are much faster than  
<sup>585</sup> the mismatched trials ( $P_{PosteriorComparison} = .75$ ). We focused on matched trials only, and  
<sup>586</sup> compared different conditions: Good () is not faster than the neutral (),  
<sup>587</sup>  $P_{PosteriorComparison} = .5$ , it was faster than the Bad condition (),  
<sup>588</sup>  $P_{PosteriorComparison} = .88$ . And the neutral condition is faster than the bad condition,  
<sup>589</sup>  $P_{PosteriorComparison} = .95$ . However, the mismatched trials are largely overlapped.

<sup>590</sup> **HDDM.** We fitted our data with HDDM, using the response-coding (also see Hu et  
<sup>591</sup> al., 2020). We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary  
<sup>592</sup> separation ( $a$ ) for each condition. We found that the shapes tagged with good person has

593 higher drift rate and higher boundary separation than shapes tagged with both neutral and  
594 bad person. Also, the shapes tagged with neutral person has a higher drift rate than  
595 shapes tagged with bad person, but not for the boundary separation. Finally, we found  
596 that shapes tagged with bad person had longer non-decision time (see figure 9)).

597 **Experiment 2: Sequential presenting**

598 Experiment 2 was conducted for two purpose: (1) to further confirm the facilitation  
599 effect of positive moral associations; (2) to test the effect of expectation of occurrence of  
600 each pair. In this experiment, after participant learned the association between labels and  
601 shapes, they were presented a label first and then a shape, they then asked to judge  
602 whether the shape matched the label or not (see (Sui, Sun, Peng, & Humphreys, 2014).  
603 Previous studies showed that when the labels presented before the shapes, participants  
604 formed expectations about the shape, and therefore a top-down process were introduced  
605 into the perceptual matching processing. If the facilitation effect of positive moral valence  
606 we found in experiment 1 was mainly drive by top-down processes, this sequential  
607 presenting paradigm may eliminate or attenuate this effect; if, however, the facilitation  
608 effect occurred because of button-up processes, then, similar facilitation effect will appear  
609 even with sequential presenting paradigm.

610 **Method.**

611 ***Participants.***

612 35 participants (17 female, age =  $21.66 \pm 3.03$ ) were recruited. 24 of them had  
613 participated in Experiment 1a (9 male, mean age = 21.9, s.d. = 2.9), and the time gap  
614 between these experiment 1a and experiment 2 is at least six weeks. The results of 1  
615 participants were excluded from analysis because of less than 60% overall accuracy,  
616 remains 34 participants (17 female, age =  $21.74 \pm 3.04$ ).

617 ***Procedure.***

In Experiment 2, the sequential presenting makes the matching task much easier than experiment 1. To avoid ceiling effect on behavioral data, we did a few pilot experiments to get optimal parameters, i.e., the conditions under which participant have similar accuracy as in Experiment 1 (around 70 ~ 80% accuracy). In the final procedure, the label (good person, bad person, or neutral person) was presented for 50 ms and then masked by a scrambled image for 200 ms. A geometric shape followed the scrambled mask for 50 ms in a noisy background (which was produced by first decomposing a square with  $\frac{3}{4}$  gray area and  $\frac{1}{4}$  white area to small squares with a size of  $2 \times 2$  pixels and then re-combine these small pieces randomly), instead of pure gray background in Experiment 1. After that, a blank screen was presented 1100 ms, during which participants should press a button to indicate the label and the shape match the original association or not. Feedback was given, as in study 1. The next trial then started after 700 ~ 1100 ms blank. Other aspects of study 2 were identical to study 1.

### ***Data analysis.***

Data was analyzed as in study 1a.

### **Results.**

#### ***NHST.***

Figure 10 shows  $d$  prime and reaction times of experiment 2. Less than 0.2% correct trials with less than 200ms reaction times were excluded.

#### *d prime.*

There was evidence for the main effect of valence,  $F(1.83, 60.36) = 14.41$ ,  $MSE = 0.23$ ,  $p < .001$ ,  $\eta^2_G = .066$ . Paired t test showed that the Good-Person condition ( $2.79 \pm 0.17$ ) was with greater  $d$  prime than Netural condition ( $2.21 \pm 0.16$ ,  $t(33) = 4.723$ ,  $p = 0.001$ ) and Bad-person condition ( $2.41 \pm 0.14$ ),  $t(33) = 4.067$ ,  $p = 0.008$ ). There was no-significant difference between Neutral-person and Bad-person conditition,  $t(33) = -1.802$ ,  $p = 0.185$ .

644 *Reaction time.*

645 The results of reaction times of matchness trials showed similar pattern as the  $d$   
 646 prime data.

647 We found interaction between Matchness and Valence ( $F(1.99, 65.70) = 9.53$ ,  
 648  $MSE = 605.36$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .017$ ) and then analyzed the matched trials and  
 649 mismatched trials separately, as in experiment 1a. For matched trials, we found the effect  
 650 of valence  $F(1.99, 65.76) = 10.57$ ,  $MSE = 1,192.65$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .067$ . Post-hoc  $t$ -tests  
 651 revealed that shapes associated with Good Person ( $548 \pm 9.4$ ) were responded faster than  
 652 Neutral-Person ( $582 \pm 10.9$ ), ( $t(33) = -3.95$ ,  $p = 0.0011$ ) and Bad Person ( $582 \pm 10.2$ ),  
 653  $t(33) = -3.9$ ,  $p = 0.0013$ ). While there was no significant differences between Neutral and  
 654 Bad-Person condition ( $t(33) = -0.01$ ,  $p = 0.999$ ). For non-matched trials, there was no  
 655 significant effect of Valence ( $F(1.99, 65.83) = 0.17$ ,  $MSE = 489.80$ ,  $p = .843$ ,  $\hat{\eta}_G^2 = .001$ ).

656 **BGLMM.**

657 *Signal detection theory analysis of accuracy.*

658 We fitted a Bayesian hierarchical GLM for SDT. The results showed that when the  
 659 shapes were tagged with labels with different moral valence, the sensitivity ( $d'$ ) and criteria  
 660 ( $c$ ) were both influence. For the  $d'$ , we found that the shapes tagged with morally good  
 661 person ( $2.46$ , 95% CI[ $2.21$   $2.72$ ]) is greater than shapes tagged with moral bad ( $2.07$ , 95%  
 662 CI[ $1.83$   $2.32$ ]),  $P_{PosteriorComparison} = 1$ . Shape tagged with morally good person is also  
 663 greater than shapes tagged with neutral person ( $2.23$ , 95% CI[ $1.95$   $2.49$ ]),  
 664  $P_{PosteriorComparison} = 0.97$ . Also, the shapes tagged with neutral person is greater than  
 665 shapes tagged with morally bad person,  $P_{PosteriorComparison} = 0.92$ .

666 Interesting, we also found the criteria for three conditions also differ, the shapes  
 667 tagged with good person has the highest criteria ( $-1.01$ , [- $1.14$   $-0.88$ ]), followed by shapes  
 668 tagged with neutral person( $1.06$ , [- $1.21$   $-0.92$ ]), and then the shapes tagged with bad  
 669 person( $-1.11$ , [- $1.25$   $-0.97$ ]). However, pair-wise comparison showed that only showed strong

670 evidence for the difference between good and bad conditions.

671 *Reaction times.*

672 We fitted a Bayesian hierarchical GLM for RTs, with a log-normal distribution as the  
673 link function. We used the posterior distribution of the regression coefficient to make  
674 statistical inferences. As in previous studies, the matched conditions are much faster than  
675 the mismatched trials ( $P_{PosteriorComparison} = .75$ ). We focused on matched trials only, and  
676 compared different conditions: Good () is not faster than the neutral (),  
677  $P_{PosteriorComparison} = .5$ , it was faster than the Bad condition (),  
678  $P_{PosteriorComparison} = .88$ . And the neutral condition is faster than the bad condition,  
679  $P_{PosteriorComparison} = .95$ . However, the mismatched trials are largely overlapped.

680 **HDDM.** We fitted our data with HDDM, using the response-coding (also see Hu et  
681 al., 2020). We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary  
682 separation ( $a$ ) for each condition. We found that the shapes tagged with good person has  
683 higher drift rate and higher boundary separation than shapes tagged with both neutral and  
684 bad person. Also, the shapes tagged with neutral person has a higher drift rate than  
685 shapes tagged with bad person, but not for the boundary separation. Finally, we found  
686 that shapes tagged with bad person had longer non-decision time (see figure  
687 @ref(fig:plot-exp1c -HDDM))).

## 688 Discussion

689 In this experiment, we repeated the results pattern that the positive moral valenced  
690 stimuli has an advantage over the neutral or the negative valence association. Moreover,  
691 with a cross-task analysis, we didn't find evidence that the experiment task interacted  
692 with moral valence, suggesting that the effect might not be effect by experiment task.  
693 These findings suggested that the facilitation effect of positive moral valence is robust and  
694 not affected by task. This robust effect detected by the associative learning is unexpected.

695 **Experiment 6a: EEG study 1**

696 Experiment 6a was conducted to study the neural correlates of the positive  
697 prioritization effect. The behavioral paradigm is same as experiment 2.

698 **Method.**

699 ***Participants.***

700 24 college students (8 female, age =  $22.88 \pm 2.79$ ) participated the current study, all  
701 of them were from Tsinghua University in 2014. Informed consent was obtained from all  
702 participants prior to the experiment according to procedures approved by a local ethics  
703 committee. No participant was excluded from behavioral analysis.

704 **Experimental design.** The experimental design of this experiment is same as  
705 experiment 2: a  $3 \times 2$  within-subject design with moral valence (good, neutral and bad  
706 associations) and matchness between shape and label (match vs. mismatch for the personal  
707 association) as within-subject variables.

708 ***Stimuli.***

709 Three geometric shapes (triangle, square and circle, each  $4.6^\circ \times 4.6^\circ$  of visual angle)  
710 were presented at the center of screen for 50 ms after 500ms of fixation ( $0.8^\circ \times 0.8^\circ$  of  
711 visual angle). The association of the three shapes to bad person (“ , HuaiRen”), good  
712 person (“ , HaoRen”) or ordinary person (“ , ChangRen”) was counterbalanced across  
713 participants. The words bad person, good person or ordinary person ( $3.6^\circ \times 1.6^\circ$ ) was also  
714 displayed at the center fo the screen. Participants had to judge whether the pairings of  
715 label and shape matched (e.g., Does the circle represent a bad person?). The experiment  
716 was run on a PC using E-prime software (version 2.0). These stimuli were displayed on a  
717 22-in CRT monitor ( $1024 \times 768$  at 100Hz). We used backward masking to avoid  
718 over-processing of the moral words, in which a scrambled picture were presented for 900 ms  
719 after the label. Also, to avoid the ceiling effect on accuracy, shapes were presented on a

720 noisy background based on our pilot studies. The noisy images were made by scrambling a  
721 picture of 3/4 gray and 1/4 white at resolution of 2 × 2 pixel.

722 ***Procedure.***

723 The procedure was similar to Experiment 2. Participants finished 9 blocks of trial,  
724 each with 120 trials. In total, participants finished 180 trials for each combination of  
725 condition.

726 As in experiment 2 (Sui, He, & Humphreys, 2012), subjects first learned the  
727 associations between labels and shapes and then completed a shape-label matching task  
728 (e.g., good person-triangle). In each trial of the matching task, a fixation were first  
729 presented for 500 ms, followed by a 50 ms label; then, a scrambled picture presented 900  
730 ms. After the backward mask, the shape were presented on a noisy background for 50ms.  
731 Participant have to response in 1000ms after the presentation of the shape, and finally, a  
732 feedback screen was presented for 500 ms (see figure 1). The inter-trial interval (ITI) were  
733 randomly varied at the range of 1000 ~ 1400 ms.

734 All the stimuli were presented on a gray background (RGB: 127, 127, 127). E-primed  
735 2.0 was used to present stimuli and collect behavioral results. Data were collected and  
736 analyzed when accuracy performance in total reached 60%.

737 **Data Analysis.** Data was analyzed as in experiment 1a.

738 **Results.**

739 **NHST.**

740 Only the behavioral results were reported here. Figure 13 shows  $d$  prime and reaction  
741 times of experiment 6a.

742  $d$  prime.

743 We conducted repeated measures ANOVA, with moral valence as independent  
744 variable. The results revealed the main effect of valence ( $F(1.74, 40.05) = 3.76$ ,

745  $MSE = 0.10$ ,  $p = .037$ ,  $\hat{\eta}_G^2 = .021$ ). Post-hoc analysis revealed that shapes link with Good  
 746 person (mean = 3.13, SE = 0.109) is greater than Neutral condition (mean = 2.88, SE =  
 747 0.14),  $t = 2.916$ ,  $df = 24$ ,  $p = 0.02$ , p-value adjusted by Tukey method, but the  $d$  prime  
 748 between Good and bad (mean = 3.03, SE = 0.142) ( $t = 1.512$ ,  $df = 24$ ,  $p = 0.3034$ , p-value  
 749 adjusted by Tukey method), bad and neutral ( $t = 1.599$ ,  $df = 24$ ,  $p = 0.2655$ , p-value  
 750 adjusted by Tukey method) were not significant.

751 *Reaction times.*

752 The results of reaction times of matchness trials showed similar pattern as the  $d$   
 753 prime data.

754 We found intercation between Matchness and Valence ( $F(1.97, 45.20) = 20.45$ ,  
 755  $MSE = 450.47$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .021$ ) and then analyzed the matched trials and  
 756 mismatched trials separately, as in experiment 2. For matched trials, we found the effect of  
 757 valence  $F(1.97, 45.25) = 32.37$ ,  $MSE = 522.42$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .078$ . For non-matched  
 758 trials, there was no significant effect of Valence ( $F(1.77, 40.67) = 0.35$ ,  $MSE = 242.15$ ,  
 759  $p = .679$ ,  $\hat{\eta}_G^2 = .000$ ). Post-hoc  $t$ -tests revealed that shapes associated with Good Person  
 760 (mean = 550, SE = 13.8) were responded faster than Neutral-Person (501, SE = 14.7),  
 761 ( $t(24) = -5.171$ ,  $p = 0.0001$ ) and Bad Person (523, SE = 16.3),  $t(24) = -8.137$ ,  $p <$   
 762 0.0001., and Neutral is faster than Bad-Person condition ( $t(32) = -3.282$ ,  $p = 0.0085$ ).

763 **BGLM.**

764 *Signal detection theory analysis of accuracy.*

765 *Reaction time.*

## 766 Part 2: interaction between valence and identity

767 In this part, we report two experiments that aimed at testing whether the moral  
 768 valence effect found in the previous experiment can be modulated by the self-referential  
 769 processing.

770 **Experiment 3a**

771 To examine the modulation effect of positive valence was an intrinsic, self-referential  
772 process, we designed study 3. In this study, moral valence was assigned to both self and a  
773 stranger. We hypothesized that the modulation effect of moral valence will be stronger for  
774 the self than for a stranger.

775 **Method.**

776 **Participants.**

777 38 college students (15 female, age =  $21.92 \pm 2.16$ ) participated in experiment 3a.

778 All of them were right-handed, and all had normal or corrected-to-normal vision. Informed  
779 consent was obtained from all participants prior to the experiment according to procedures  
780 approved by a local ethics committee. One female and one male student did not finish the  
781 experiment, and 1 participants' data were excluded from analysis because less than 60%  
782 overall accuracy, remains 35 participants (13 female, age =  $22.11 \pm 2.13$ ).

783 **Design.**

784 Study 3a combined moral valence with self-relevance, hence the experiment has a  $2 \times$   
785  $3 \times 2$  within-subject design. The first variable was self-relevance, include two levels:  
786 self-relevance vs. stranger-relevance; the second variable was moral valence, include good,  
787 neutral and bad; the third variable was the matching between shape and label: match  
788 vs. nonmatch.

789 **Stimuli.**

790 The stimuli used in study 3a share the same parameters with experiment 1 & 2. The  
791 differences was that we used six shapes: triangle, square, circle, trapezoid, diamond,  
792 regular pentagon, and six labels: good self, neutral self, bad self, good person, bad person,  
793 and neutral person. To match the concreteness of the label, we asked participant to chosen  
794 an unfamiliar name of their own gender to be the stranger.

**795      *Procedure.***

796      After being fully explained and signed the informed consent, participants were  
797      instructed to chose a name that can represent a stranger with same gender as the  
798      participant themselves, from a common Chinese name pool. Before experiment, the  
799      experimenter explained the meaning of each label to participants. For example, the “good  
800      self” mean the morally good side of themselves, them could imagine the moment when they  
801      do something’s morally applauded, “bad self” means the morally bad side of themselves,  
802      they could also imagine the moment when they doing something morally wrong, and  
803      “neutral self” means the aspect of self that does not related to morality, they could imagine  
804      the moment when they doing something irrelevant to morality. In the same sense, the  
805      “good other”, “bad other”, and “neutral other” means the three different aspects of the  
806      stranger, whose name was chosen before the experiment. Then, the experiment proceeded  
807      as study 1a. Each participant finished 6 blocks, each have 120 trials. The sequence of trials  
808      was pseudo-randomized so that there are 10 matched trials for each condition and 10  
809      non-matched trials for each condition (good self, neutral self, bad self, good other, neutral  
810      other, bad other) for each block.

**811      *Data Analysis.***

812      Data analysis followed strategies described in the general method section. Reaction  
813      times and  $d$  prime data were analyzed as in study 1 and study 2, except that one more  
814      within-subject variable (i.e., self-relevance) was included in the analysis.

**815      *Results.*****816      *NHST.***

817      Figure 15 shows  $d$  prime and reaction times of experiment 3a. Less than 5% correct  
818      trials with less than 200ms reaction times were excluded.

**819      *d prime.***

820 There was evidence for the main effect of valence,  $F(1.89, 64.37) = 11.09$ ,

821  $MSE = 0.23$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .039$ , and main effect of self-relevance,  $F(1, 34) = 3.22$ ,

822  $MSE = 0.54$ ,  $p = .082$ ,  $\hat{\eta}_G^2 = .015$ , as well as the interaction,  $F(1.79, 60.79) = 3.39$ ,

823  $MSE = 0.43$ ,  $p = .045$ ,  $\hat{\eta}_G^2 = .022$ .

824 We then conducted separated ANOVA for self-referential and other-referential trials.

825 The valence effect was shown for the self-referential conditions,  $F(1.65, 56.25) = 13.98$ ,

826  $MSE = 0.31$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .119$ . Post-hoc test revealed that the Good-Self condition

827 ( $1.97 \pm 0.14$ ) was with greater  $d$  prime than Neutral condition ( $1.41 \pm 0.12$ ,  $t(34) = 4.505$ ,

828  $p = 0.0002$ ), and Bad-self condition ( $1.43 \pm 0.102$ ),  $t(34) = 3.856$ ,  $p = 0.0014$ . There was

829 difference between neutral and bad condition,  $t(34) = -0.238$ ,  $p = 0.9694$ . However, no

830 effect of valence was found for the other-referential condition  $F(1.98, 67.36) = 0.38$ ,

831  $MSE = 0.35$ ,  $p = .681$ ,  $\hat{\eta}_G^2 = .004$ .

### 832 Reaction time.

833 We found interaction between Matchness and Valence ( $F(1.98, 67.44) = 26.29$ ,

834  $MSE = 730.09$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .025$ ) and then analyzed the matched trials and nonmatch

835 trials separately, as in previous experiments.

836 For the match trials, we found that the interaction between identity and valence,

837  $F(1.72, 58.61) = 3.89$ ,  $MSE = 2,750.19$ ,  $p = .032$ ,  $\hat{\eta}_G^2 = .019$ , as well as the main effect of

838 valence  $F(1.98, 67.34) = 35.76$ ,  $MSE = 1,127.25$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .079$ , but not the effect of

839 identity  $F(1, 34) = 0.20$ ,  $MSE = 3,507.14$ ,  $p = .660$ ,  $\hat{\eta}_G^2 = .001$ . As for the  $d$  prime, we

840 separated analyzed the self-referential and other-referential trials. For the Self-referential

841 trials, we found the main effect of valence,  $F(1.80, 61.09) = 30.39$ ,  $MSE = 1,584.53$ ,

842  $p < .001$ ,  $\hat{\eta}_G^2 = .159$ ; for the other-referential trials, the effect of valence is weaker,

843  $F(1.86, 63.08) = 2.85$ ,  $MSE = 2,224.30$ ,  $p = .069$ ,  $\hat{\eta}_G^2 = .024$ . We then focused on the self

844 conditions: the good-self condition ( $713 \pm 12$ ) is faster than neutral- ( $776 \pm 11.8$ ),  $t(34) =$

845  $-7.396$ ,  $p < .0001$ , and bad-self ( $772 \pm 10.1$ ) conditions,  $t(34) = -5.66$ ,  $p < .0001$ . But

846 there is not difference between neutral- and bad-self conditions,  $t(34) = 0.481$ ,  $p = 0.881$ .

847 For the nonmatch trials, we didn't found any strong effect: identity,  $F(1, 34) = 3.43$ ,

848  $MSE = 660.02$ ,  $p = .073$ ,  $\hat{\eta}_G^2 = .004$ , valence  $F(1.89, 64.33) = 0.40$ ,  $MSE = 444.10$ ,

849  $p = .661$ ,  $\hat{\eta}_G^2 = .001$ , or interaction between the two  $F(1.94, 66.02) = 2.42$ ,  $MSE = 817.35$ ,

850  $p = .099$ ,  $\hat{\eta}_G^2 = .007$ .

851 **BGLM.**

852 *Signal detection theory analysis of accuracy.*

853 We found that the  $d$  prime is greater when shapes were associated with good self  
 854 condition than with neutral self or bad self, but shapes associated with bad self and neutral  
 855 self didn't show differences. Comparing the self vs other under three condition revealed  
 856 that shapes associated with good self is greater than with good other, but with a weak  
 857 evidence. In contrast, for both neutral and bad valence condition, shapes associated with  
 858 other had greater  $d$  prime than with self.

859 *Reaction time.*

860 In reaction times, we found that same trends in the match trials as in the RT: while  
 861 the shapes associated with good self was greater than with good other ( $\log \text{mean diff} =$   
 862  $-0.02858$ ,  $95\% \text{HPD}[-0.070898, 0.0154]$ ), the direction is reversed for neutral and negative  
 863 condition. see Figure 16

864 **HDDM.** We fitted our data with HDDM, using the response-coding (also see Hu et  
 865 al., 2020). We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary  
 866 separation ( $a$ ) for each condition. We found that the shapes tagged with good person has  
 867 higher drift rate and higher boundary separation than shapes tagged with both neutral and  
 868 bad person. Also, the shapes tagged with neutral person has a higher drift rate than  
 869 shapes tagged with bad person, but not for the boundary separation. Finally, we found  
 870 that shapes tagged with bad person had longer non-decision time (see figure 17)).

871    **Experiment 3b**

872       In study 3a, participants had to remember 6 pairs of association, which cause high  
873 cognitive load during the whole experiment. To eliminate the influence of cognitive load, we  
874 conducted study 3b, in which participant learn three aspect of self and stranger separately  
875 in to consecutive task. We hypothesize that we will replicate the pattern of study 3a, i.e.,  
876 the effect of moral valence only occurs for self-relevant conditions. #### Method

877       **Participants.**

878       Study 3b were finished in 2017, at that time we have calculated that the effect size  
879 (Cohen's  $d$ ) of good-person (or good-self) vs. bad-person (or bad-other) was between  $0.47 \sim$   
880 0.53, based on the data from Tsinghua community in study 1a, 1b, 2, 3a, 4a, and 4b. Based  
881 on this effect size, we estimated that 54 participants would allow we to detect the effect  
882 size of Cohen's  $= 0.5$  with 95% power and alpha = 0.05, using G\*power 3.192 (Faul,  
883 Erdfelder, Buchner, & Lang, 2009). Therefore, we planned to stop after we arrived this  
884 number. During the data collected at Wenzhou University, 61 participants (45 females; 19  
885 to 25 years of age, age =  $20.42 \pm 1.77$ ) came to the testing room and we tested all of them  
886 during a single day. All participants were right-handed, and all had normal or  
887 corrected-to-normal vision. Informed consent was obtained from all participants prior to  
888 the experiment according to procedures approved by a local ethics committee. 4  
889 participants' data were excluded from analysis because their over all accuracy was lower  
890 than 60%, 1 more participant was excluded because of zero hit rate for one condition,  
891 leaving 56 participants (43 females; 19 to 25 years old, age =  $20.27 \pm 1.60$ ).

892       **Design.**

893       Study 3b has the same experimental design as 3a, with a  $2 \times 3 \times 2$  within-subject  
894 design. The first variable was self-relevance, include two levels: self-relevant  
895 vs. stranger-relevant; the second variable was moral valence, include good, neutral and bad;  
896 the third variable was the matching between shape and label: match vs. mismatch.

897 Stimuli. The stimuli used in study 3b share the same parameters with experiment 3a. 6  
898 shapes were included (triangle, square, circle, trapezoid, diamond, regular pentagon), as  
899 well as 6 labels, but the labels changed to “good self”, “neutral self”, “bad self”, “good  
900 him/her”, bad him/her”, “neutral him/her”, the stranger’s label is consistent with  
901 participants’ gender. Same as study 3a, we asked participant to chosen an unfamiliar name  
902 of their own gender to be the stranger before showing them the relationship. Note, because  
903 of implementing error, the personal distance data did not collect for this experiment.

904 ***Stimuli.***

905 The stimuli used in study 3b is the same as in experiment 3a.

906 ***Procedure.***

907 In this experiment, participants finished two matching tasks, i.e., self-matching task,  
908 and other-matching task. In the self-matching task, participants first associate the three  
909 aspects of self to three different shapes, and then perform the matching task. In the  
910 other-matching task, participants first associate the three aspects of the stranger to three  
911 different shapes, and then perform the matching task. The order of self-task and other-task  
912 are counter-balanced among participants. Different from experiment 3a, after presenting  
913 the stimuli pair for 100ms, participant has 1900 ms to response, and they feedback with  
914 both accuracy and reaction time. As in study 3a, before each task, the instruction showed  
915 the meaning of each label to participants. The self-matching task and other-matching task  
916 were randomized between participants. Each participant finished 6 blocks, each have 120  
917 trials.

918 ***Data Analysis.***

919 Same as experiment 3a.

920 ***Results.***

921 ***NHST.***

922       Figure 18 shows  $d$  prime and reaction times of experiment 3b. Less than 5% correct

923 trials with less than 200ms reaction times were excluded.

924       *d prime.*

925       There was no evidence for the main effect of valence,  $F(1.92, 105.43) = 1.90$ ,

926        $MSE = 0.33$ ,  $p = .157$ ,  $\hat{\eta}_G^2 = .005$ , but we found a main effect of self-relevance,

927        $F(1, 55) = 4.65$ ,  $MSE = 0.89$ ,  $p = .035$ ,  $\hat{\eta}_G^2 = .017$ , as well as the interaction,

928        $F(1.90, 104.36) = 5.58$ ,  $MSE = 0.26$ ,  $p = .006$ ,  $\hat{\eta}_G^2 = .011$ .

929       We then conducted separated ANOVA for self-referential and other-referential trials.

930       The valence effect was shown for the self-referential conditions,  $F(1.75, 96.42) = 6.73$ ,

931        $MSE = 0.30$ ,  $p = .003$ ,  $\hat{\eta}_G^2 = .037$ . Post-hoc test revealed that the Good-Self condition

932        $(2.15 \pm 0.12)$  was with greater  $d$  prime than Neutral condition  $(1.83 \pm 0.12$ ,  $t(34) = 3.36$ ,

933        $p = 0.0031$ ), and Bad-self condition  $(1.87 \pm 0.12)$ ,  $t(34) = 2.955$ ,  $p = 0.01$ . There was

934       difference between neutral and bad condition,  $t(34) = -0.039$ ,  $p = 0.914$ . However, no effect

935       of valence was found for the other-referential condition  $F(1.93, 105.97) = 0.61$ ,

936        $MSE = 0.31$ ,  $p = .539$ ,  $\hat{\eta}_G^2 = .002$ .

937       *Reaction time.*

938       We found interaction between Matchness and Valence ( $F(1.86, 102.47) = 15.44$ ,

939        $MSE = 3,112.78$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .006$ ) and then analyzed the matched trials and

940       nonmatch trials separately, as in previous experiments.

941       For the match trials, we found that the interaction between identity and valence,

942        $F(1.67, 92.11) = 6.14$ ,  $MSE = 6,472.48$ ,  $p = .005$ ,  $\hat{\eta}_G^2 = .009$ , as well as the main effect of

943       valence  $F(1.88, 103.65) = 24.25$ ,  $MSE = 5,994.25$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .038$ , but not the effect

944       of identity  $F(1, 55) = 48.49$ ,  $MSE = 25,892.59$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .153$ . As for the  $d$  prime,

945       we separated analyzed the self-referential and other-referential trials. For the

946       Self-referential trials, we found the main effect of valence,  $F(1.66, 91.38) = 23.98$ ,

947        $MSE = 6,965.61$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .100$ ; for the other-referential trials, the effect of valence

is weaker,  $F(1.89, 103.94) = 5.96$ ,  $MSE = 5,589.90$ ,  $p = .004$ ,  $\hat{\eta}_G^2 = .014$ . We then focused on the self conditions: the good-self condition ( $713 \pm 12$ ) is faster than neutral- ( $776 \pm 11.8$ ),  $t(34) = -7.396$ ,  $p < .0001$ , and bad-self ( $772 \pm 10.1$ ) conditions,  $t(34) = -5.66$ ,  $p < .0001$ . But there is not difference between neutral- and bad-self conditions,  $t(34) = 0.481$ ,  $p = 0.881$ .

For the nonmatch trials, we didn't found any strong effect: identity,  $F(1, 55) = 10.31$ ,  $MSE = 24,590.52$ ,  $p = .002$ ,  $\hat{\eta}_G^2 = .035$ , valence  $F(1.98, 108.63) = 20.57$ ,  $MSE = 2,847.51$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .016$ , or interaction between the two  $F(1.93, 106.25) = 35.51$ ,  $MSE = 1,939.88$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .019$ .

### 957 **BGLM.**

#### 958 *Signal detection theory analysis of accuracy.*

959 We found that the  $d$  prime is greater when shapes were associated with good self  
 960 condition than with neutral self or bad self, but shapes associated with bad self and neutral  
 961 self didn't show differences. comparing the self vs other under three condition revealed that  
 962 shapes associated with good self is greater than with good other, but with a weak evidence.  
 963 In contrast, for both neutral and bad valence condition, shapes associated with other had  
 964 greater  $d$  prime than with self.

#### 965 *Reaction time.*

966 In reaction times, we found that same trends in the match trials as in the RT: while  
 967 the shapes associated with good self was greater than with good other (log mean diff =  
 968  $-0.02858$ , 95%HPD[-0.070898, 0.0154]), the direction is reversed for neutral and negative  
 969 condition. see Figure 19

970 **HDDM.** We fitted our data with HDDM, using the response-coding (also see Hu et  
 971 al., 2020). We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary  
 972 separation ( $a$ ) for each condition. We found that, similar to experiment 3a, the shapes  
 973 tagged with good person has higher drift rate and higher boundary separation than shapes

974 tagged with both neutral and bad person, but only for the self-referential condition. Also,  
975 the shapes tagged with neutral person has a higher drift rate than shapes tagged with bad  
976 person, but not for the boundary separation, and this effect also exist only for the  
977 self-referential condition.

978 Interestingly, we found that in both self-referential and other-referential conditions,  
979 the shapes associated bad valence have higher drift rate and higher boundary separation.  
980 which might suggest that the shape associated with bad stimuli might be prioritized in the  
981 non-match trials (see figure 20)).

## 982 **Experiment 6b**

983 Experiment 6b was conducted to study the neural correlates of the prioritization  
984 effect of positive self, i.e., the neural underlying of the behavioral effect found int  
985 experiment 3a. However, as in experiment 6a, the procedure of this experiment was  
986 modified to adopted to ERP experiment.

### 987 **Method.**

#### 988 ***Participants.***

989 23 college students (8 female, age =  $22.86 \pm 2.47$ ) participated the current study, all  
990 of them were recruited from Tsinghua University in 2016. Informed consent was obtained  
991 from all participants prior to the experiment according to procedures approved by a local  
992 ethics committee. For day 1's data, 1 participant was excluded from the current analysis  
993 because of lower than 60% overall accuracy, remaining 22 participants (8 female, age =  
994  $22.76 \pm 2.49$ ). For day 2's data, one participant dropped out, leaving 22 participants (9  
995 female, age =  $23.05 \pm 2.46$ ), all of them has overall accuracy higher than 60%.

#### 996 ***Design.***

997 The experimental design of this experiment is same as experiment 3: a  $2 \times 3 \times 2$   
998 within-subject design with self-relevance (self-relevant vs. other-relevant), moral valence

999 (good, neutral, and bad) and matchness between shape and label (match vs. mismatch) as  
1000 within-subject variables.

1001 ***Stimuli.***

1002 As in experiment 3a, 6 shapes were included (triangle, square, circle, trapezoid,  
1003 diamond, regular pentagon), as well as 6 labels (good self, neutral self, bad self, good  
1004 person, bad person, neutral person). To match the concreteness of the label, we asked  
1005 participant to chosen an unfamiliar name of their own gender to be the stranger.

1006 ***Procedure.***

1007 The procedure was similar to Experiment 2 and 6a. Subjects first learned the  
1008 associations between labels and shapes and then completed a shape-label matching task. In  
1009 each trial of the matching task, a fixation were first presented for 500 ms, followed by a 50  
1010 ms label; then, a scrambled picture presented 900 ms. After the backward mask, the shape  
1011 were presented on a noisy background for 50ms. Participant have to response in 1000ms  
1012 after the presentation of the shape, and finally, a feedback screen was presented for 500 ms.  
1013 The inter-trial interval (ITI) were randomly varied at the range of 1000 ~ 1400 ms.

1014 All the stimuli were presented on a gray background (RGB: 127, 127, 127). E-primed  
1015 2.0 was used to present stimuli and collect behavioral results. Data were collected and  
1016 analyzed when accuracy performance in total reached 60%.

1017 Because learning 6 associations was more difficult than 3 associations and participant  
1018 might have low accuracy (see experiment 3a), the current study had extended to a two-day  
1019 paradigm to maximizing the accurate trials that can be used in EEG data. At the first day,  
1020 participants learnt the associations and finished 9 blocks of the matching task, each had  
1021 120 trials, without EEG recording. That is, each condition has 90 trials.

1022 Participants came back to lab at the second day and finish the same task again, with  
1023 EEG recorded. Before the EEG experiment, each participant finished a practice session  
1024 again, if their accuracy is equal or higher than 85%, they start the experiment (one

1025 participant used lower threshold 75%). Each participant finished 18 blocks, each has 90  
 1026 trials. One participant finished additional 6 blocks because of high error rate at the  
 1027 beginning, another two participant finished addition 3 blocks because of the technique  
 1028 failure in recording the EEG data. To increase the number of trials that can be used for  
 1029 EEG data analysis, matched trials has twice number as mismatched trials, therefore, for  
 1030 matched trials each participants finished 180 trials for each condition, for mismatched  
 1031 trials, each conditions has 90 trials.

1032 ***Data Analysis.***

1033 Same as experiment 3a.

1034 **Results of Day 1.**

1035 ***NHST.***

1036 Figure 21 shows *d* prime and reaction times of experiment 3b. Less than 5% correct  
 1037 trials with less than 200ms reaction times were excluded.

1038 *d prime.*

1039 There was no evidence for the main effect of valence,  $F(1.91, 40.20) = 11.98$ ,  
 1040  $MSE = 0.15$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .040$ , but we found a main effect of self-relevance,  
 1041  $F(1, 21) = 1.21$ ,  $MSE = 0.20$ ,  $p = .284$ ,  $\hat{\eta}_G^2 = .003$ , as well as the interaction,  
 1042  $F(1.28, 26.90) = 12.88$ ,  $MSE = 0.21$ ,  $p = .001$ ,  $\hat{\eta}_G^2 = .041$ .

1043 We then conducted separated ANOVA for self-referential and other-referential trials.  
 1044 The valence effect was shown for the self-referential conditions,  $F(1.73, 36.42) = 29.31$ ,  
 1045  $MSE = 0.14$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .147$ . Post-hoc test revealed that the Good-Self condition  
 1046 ( $2.15 \pm 0.12$ ) was with greater *d* prime than Neutral condition ( $1.83 \pm 0.12$ ,  $t(34) = 3.36$ ,  
 1047  $p = 0.0031$ ), and Bad-self condition ( $1.87 \pm 0.12$ ),  $t(34) = 2.955$ ,  $p = 0.01$ . There was  
 1048 difference between neutral and bad condition,  $t(34) = -0.039$ ,  $p = 0.914$ . However, no effect  
 1049 of valence was found for the other-referential condition  $F(1.75, 36.72) = 0.00$ ,  $MSE = 0.18$ ,  
 1050  $p = .999$ ,  $\hat{\eta}_G^2 = .000$ .

1051        *Reaction time.*

1052        We found interaction between Matchness and Valence ( $F(1.79, 37.63) = 4.07$ ,

1053         $MSE = 704.90, p = .029, \hat{\eta}_G^2 = .003$ ) and then analyzed the matched trials and nonmatch

1054        trials separately, as in previous experiments.

1055        For the match trials, we found that the interaction between identity and valence,

1056         $F(1.72, 36.16) = 4.55, MSE = 1,560.90, p = .022, \hat{\eta}_G^2 = .015$ , as well as the main effect of

1057        valence  $F(1.93, 40.55) = 9.83, MSE = 1,951.84, p < .001, \hat{\eta}_G^2 = .044$ , but not the effect of

1058        identity  $F(1, 21) = 4.87, MSE = 2,032.05, p = .039, \hat{\eta}_G^2 = .012$ . As for the  $d$  prime, we

1059        separated analyzed the self-referential and other-referential trials. For the Self-referential

1060        trials, we found the main effect of valence,  $F(1.92, 40.38) = 14.48, MSE = 1,647.20$ ,

1061         $p < .001, \hat{\eta}_G^2 = .112$ ; for the other-referential trials, the effect of valence is weaker,

1062         $F(1.79, 37.50) = 1.04, MSE = 1,842.07, p = .356, \hat{\eta}_G^2 = .008$ . We then focused on the self

1063        conditions: the good-self condition ( $713 \pm 12$ ) is faster than neutral- ( $776 \pm 11.8$ ),  $t(34) =$

1064         $-7.396, p < .0001$ , and bad-self ( $772 \pm 10.1$ ) conditions,  $t(34) = -5.66, p < .0001$ . But

1065        there is not difference between neutral- and bad-self conditions,  $t(34) = 0.481, p = 0.881$ .

1066        For the nonmatch trials, we didn't found any strong effect: identity,  $F(1, 21) = 2.76$ ,

1067         $MSE = 1,718.93, p = .112, \hat{\eta}_G^2 = .006$ , valence  $F(1.61, 33.77) = 3.81, MSE = 1,532.21$ ,

1068         $p = .041, \hat{\eta}_G^2 = .012$ , or interaction between the two  $F(1.90, 39.97) = 2.23, MSE = 720.80$ ,

1069         $p = .123, \hat{\eta}_G^2 = .004$ .

1070        **BGLM.**

1071        *Signal detection theory analysis of accuracy.*

1072        We found that the  $d$  prime is greater when shapes were associated with good self

1073        condition than with neutral self or bad self, but shapes associated with bad self and neutral

1074        self didn't show differences. comparing the self vs other under three condition revealed that

1075        shapes associated with good self is greater than with good other, but with a weak evidence.

1076        In contrast, for both neutral and bad valence condition, shapes associated with other had

1077 greater  $d$  prime than with self.

1078 *Reaction time.*

1079 In reaction times, we found that same trends in the match trials as in the RT: while  
1080 the shapes associated with good self was greater than with good other (log mean diff =  
1081 -0.02858, 95%HPD[-0.070898, 0.0154]), the direction is reversed for neutral and negative  
1082 condition. see Figure 22

1083 **HDDM.** We fitted our data with HDDM, using the response-coding (also see Hu et  
1084 al., 2020). We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary  
1085 separation ( $a$ ) for each condition. We found that, similar to experiment 3a, the shapes  
1086 tagged with good person has higher drift rate and higher boundary separation than shapes  
1087 tagged with both neutral and bad person, but only for the self-referential condition. Also,  
1088 the shapes tagged with neutral person has a higher drift rate than shapes tagged with bad  
1089 person, but not for the boundary separation, and this effect also exist only for the  
1090 self-referential condition.

1091 Interestingly, we found that in both self-referential and other-referential conditions,  
1092 the shapes associated bad valence have higher drift rate and higher boundary separation.  
1093 which might suggest that the shape associated with bad stimuli might be prioritized in the  
1094 non-match trials (see figure 23)).

1095 **Part 3: Implicit binding between valence and identity**

1096 In this part, we reported two studies in which the moral valence or the self-referential  
1097 processing is not task-relevant. We are interested in testing whether the task-relevance will  
1098 eliminate the effect observed in previous experiment.

1099 **Experiment 4a**

1100 In part two (experiment 3a and 3b), participants learned the association between self  
1101 and moral valence directly. In Experiment 4a, we examined whether the interaction  
1102 between moral valence and identity occur even when one of the variable was irrelevant to  
1103 the task. In experiment 4a, participants learnt associations between shapes and self/other  
1104 labels, then made perceptual match judgments only about the self or other conditions  
1105 labels and shapes (cf. Sui et al. (2012)). However, we presented labels of different moral  
1106 valence in the shapes, which means that the moral valence factor become task irrelevant. If  
1107 the binding between moral good and self is intrinsic and automatic, then we will observe  
1108 that facilitating effect of moral good for self conditions, but not for other conditions.

1109 **Method.**

1110 ***Participants.***

1111 64 participants (37 female, age =  $19.70 \pm 1.22$ ) participated the current study, 32 of  
1112 them were from Tsinghua University in 2015, 32 were from Wenzhou University  
1113 participated in 2017. All participants were right-handed, and all had normal or  
1114 corrected-to-normal vision. Informed consent was obtained from all participants prior to  
1115 the experiment according to procedures approved by a local ethics committee. The data  
1116 from 5 participants from Wenzhou site were excluded from analysis because their accuracy  
1117 was close to chance ( $< 0.6$ ). The results for the remaining 59 participants (33 female, age  
1118 =  $19.78 \pm 1.20$ ) were analyzed and reported.

1119 ***Design.***

1120 As in Experiment 3, a  $2 \times 3 \times 2$  within-subject design was used. The first variable was  
1121 self-relevance (self and stranger associations); the second variable was moral valence (good,  
1122 neutral and bad associations); the third variable was the matching between shape and label  
1123 (matching vs. non-match for the personal association). However, in this the task,  
1124 participants only learn the association between two geometric shapes and two labels (self

and other), i.e., only self-relevance were related to the task. The moral valence manipulation was achieved by embedding the personal label of the labels in the geometric shapes, see below. For simplicity, the trials where shapes where paired with self and with a word of “good person” inside were shorted as good-self condition, similarly, the trials where shapes paired with the self and with a word of “bad person” inside were shorted as bad-self condition. Hence, we also have six conditions: good-self, neutral-self, bad-self, good-other, neutral-other, and bad-other.

***Stimuli.***

2 shapes were included (circle, square) and each appeared above a central fixation cross with the personal label appearing below. However, the shapes were not empty but with a two-Chinese-character word in the middle, the word was one of three labels with different moral valence: “good person”, “bad person” and “neutral person”. Before the experiment, participants learned the self/other association, and were informed to only response to the association between shapes’ configures and the labels below the fixation, but ignore the words within shapes. Besides the behavioral experiments, participants from Tsinghua community also finished questionnaires as Experiments 3, and participants from Wenzhou community finished a series of questionnaire as the other experiment finished in Wenzhou.

***Procedure.***

The procedure was similar to Experiment 1. There were 6 blocks of trial, each with 120 trials for 2017 data. Due to procedure error, the data collected in 2015 in Tsinghua community only have 60 trials for each block, i.e., 30 trials per condition.

As in study 3a, before each task, the instruction showed the meaning of each label to participants. The self-matching task and other-matching task were randomized between participants. Each participant finished 6 blocks, each have 120 trials.

***Data Analysis.***

<sub>1151</sub> Same as experiment 3a.

<sub>1152</sub> **Results.**

<sub>1153</sub> **NHST.**

<sub>1154</sub> Figure 24 shows  $d$  prime and reaction times of experiment 3a. Less than 5% correct  
<sub>1155</sub> trials with less than 200ms reaction times were excluded.

<sub>1156</sub>  $d$  prime.

<sub>1157</sub> There was no evidence for the main effect of valence,  $F(1.93, 111.66) = 0.53$ ,  
<sub>1158</sub>  $MSE = 0.12$ ,  $p = .581$ ,  $\hat{\eta}_G^2 = .000$ , but we found a main effect of self-relevance,  
<sub>1159</sub>  $F(1, 58) = 121.04$ ,  $MSE = 0.48$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .189$ , as well as the interaction,  
<sub>1160</sub>  $F(1.99, 115.20) = 4.12$ ,  $MSE = 0.14$ ,  $p = .019$ ,  $\hat{\eta}_G^2 = .004$ .

<sub>1161</sub> We then conducted separated ANOVA for self-referential and other-referential trials.  
<sub>1162</sub> The valence effect was shown for the self-referential conditions,  $F(1.95, 112.92) = 3.01$ ,  
<sub>1163</sub>  $MSE = 0.15$ ,  $p = .055$ ,  $\hat{\eta}_G^2 = .008$ . Post-hoc test revealed that the Good-Self condition  
<sub>1164</sub> ( $2.15 \pm 0.12$ ) was with greater  $d$  prime than Neutral condition ( $1.83 \pm 0.12$ ,  $t(34) = 3.36$ ,  
<sub>1165</sub>  $p = 0.0031$ ), and Bad-self condition ( $1.87 \pm 0.12$ ),  $t(34) = 2.955$ ,  $p = 0.01$ . There was  
<sub>1166</sub> difference between neutral and bad condition,  $t(34) = -0.039$ ,  $p = 0.914$ . However, no effect  
<sub>1167</sub> of valence was found for the other-referential condition  $F(1.98, 114.61) = 1.75$ ,  
<sub>1168</sub>  $MSE = 0.10$ ,  $p = .179$ ,  $\hat{\eta}_G^2 = .003$ .

<sub>1169</sub> *Reaction time.*

<sub>1170</sub> We found interaction between Matchness and Valence ( $F(1.94, 112.64) = 0.84$ ,  
<sub>1171</sub>  $MSE = 465.35$ ,  $p = .432$ ,  $\hat{\eta}_G^2 = .000$ ) and then analyzed the matched trials and nonmatch  
<sub>1172</sub> trials separately, as in previous experiments.

<sub>1173</sub> For the match trials, we found that the interaction between identity and valence,  
<sub>1174</sub>  $F(1.90, 110.18) = 4.41$ ,  $MSE = 465.91$ ,  $p = .016$ ,  $\hat{\eta}_G^2 = .003$ , as well as the main effect of  
<sub>1175</sub> valence  $F(1.98, 114.82) = 0.94$ ,  $MSE = 606.30$ ,  $p = .392$ ,  $\hat{\eta}_G^2 = .001$ , but not the effect of

identity  $F(1, 58) = 124.15$ ,  $MSE = 4,037.53$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .257$ . As for the  $d$  prime, we  
 separated analyzed the self-referential and other-referential trials. For the Self-referential  
 trials, we found the main effect of valence,  $F(1.97, 114.32) = 6.29$ ,  $MSE = 367.25$ ,  
 $p = .003$ ,  $\hat{\eta}_G^2 = .006$ ; for the other-referential trials, the effect of valence is weaker,  
 $F(1.95, 112.89) = 0.35$ ,  $MSE = 699.50$ ,  $p = .699$ ,  $\hat{\eta}_G^2 = .001$ . We then focused on the self  
 conditions: the good-self condition ( $713 \pm 12$ ) is faster than neutral- ( $776 \pm 11.8$ ),  $t(34) =$   
 $-7.396$ ,  $p < .0001$ , and bad-self ( $772 \pm 10.1$ ) conditions,  $t(34) = -5.66$ ,  $p < .0001$ . But  
 there is not difference between neutral- and bad-self conditions,  $t(34) = 0.481$ ,  $p = 0.881$ .

For the nonmatch trials, we didn't found any strong effect: identity,  $F(1, 58) = 0.16$ ,  
 $MSE = 1,547.37$ ,  $p = .692$ ,  $\hat{\eta}_G^2 = .000$ , valence  $F(1.96, 113.52) = 0.68$ ,  $MSE = 390.26$ ,  
 $p = .508$ ,  $\hat{\eta}_G^2 = .000$ , or interaction between the two  $F(1.90, 110.27) = 0.04$ ,  
 $MSE = 585.80$ ,  $p = .953$ ,  $\hat{\eta}_G^2 = .000$ .

## 1188 **BGLM.**

1189 *Signal detection theory analysis of accuracy.*

1190 We found that the  $d$  prime is greater when shapes were associated with good self  
 1191 condition than with neutral self or bad self, but shapes associated with bad self and neutral  
 1192 self didn't show differences. comparing the self vs other under three condition revealed that  
 1193 shapes associated with good self is greater than with good other, but with a weak evidence.  
 1194 In contrast, for both neutral and bad valence condition, shapes associated with other had  
 1195 greater  $d$  prime than with self.

1196 *Reaction time.*

1197 In reaction times, we found that same trends in the match trials as in the RT: while  
 1198 the shapes associated with good self was greater than with good other (log mean diff =  
 1199  $-0.02858$ , 95%HPD[-0.070898, 0.0154]), the direction is reversed for neutral and negative  
 1200 condition. see Figure 25

1201       **HDDM.** We fitted our data with HDDM, using the response-coding (also see Hu et  
1202 al., 2020). We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary  
1203 separation ( $a$ ) for each condition. We found that the shapes tagged with good person has  
1204 higher drift rate and higher boundary separation than shapes tagged with both neutral and  
1205 bad person. Also, the shapes tagged with neutral person has a higher drift rate than  
1206 shapes tagged with bad person, but not for the boundary separation. Finally, we found  
1207 that shapes tagged with bad person had longer non-decision time (see figure 26)).

1208       **Experiment 4b**

1209       In study 4b, we changed the role of valence and identity in task. In this experiment,  
1210 participants learn the association between moral valence and the made perceptual match  
1211 judgments to associations between different moral valence and shapes as in study 1-3.  
1212 Different from experiment 1 ~ 3, we made put the labels of “self/other” in the shapes so  
1213 that identity served as an task irrelevant variable. As in experiment 4b, we also  
1214 hypothesized that the intrinsic binding between morally good and self will enhance the  
1215 performance of good self condition, even identity is irrelevant to the task.

1216       **Method.**

1217       **Participants.**

1218       53 participants (39 female, age =  $20.57 \pm 1.81$ ) participated the current study, 34 of  
1219 them were from Tsinghua University in 2015, 19 were from Wenzhou University  
1220 participated in 2017. All participants were right-handed, and all had normal or  
1221 corrected-to-normal vision. Informed consent was obtained from all participants prior to  
1222 the experiment according to procedures approved by a local ethics committee. The data  
1223 from 8 participants from Wenzhou site were excluded from analysis because their accuracy  
1224 was close to chance ( $< 0.6$ ). The results for the remaining 45 participants (33 female, age  
1225 =  $20.78 \pm 1.76$ ) were analyzed and reported.

***Design.***

As in Experiment 3, a  $2 \times 3 \times 2$  within-subject design was used. The first variable was self-relevance (self and stranger associations); the second variable was moral valence (good, neutral and bad associations); the third variable was the matching between shape and label (matching vs. non-match for the personal association). However, in this task, participants only learn the association between two geometric shapes and two labels (self and other), i.e., only self-relevance were related to the task. The moral valence manipulation was achieved by embedding the personal label of the labels in the geometric shapes, see below. For simplicity, the trials where shapes where paired with self and with a word of “good person” inside were shorted as good-self condition, similarly, the trials where shapes paired with the self and with a word of “bad person” inside were shorted as bad-self condition. Hence, we also have six conditions: good-self, neutral-self, bad-self, good-other, neutral-other, and bad-other.

***Stimuli.***

2 shapes were included (circle, square) and each appeared above a central fixation cross with the personal label appearing below. However, the shapes were not empty but with a two-Chinese-character word in the middle, the word was one of three labels with different moral valence: “good person”, “bad person” and “neutral person”. Before the experiment, participants learned the self/other association, and were informed to only response to the association between shapes’ configure and the labels below the fixation, but ignore the words within shapes. Besides the behavioral experiments, participants from Tsinghua community also finished questionnaires as Experiments 3, and participants from Wenzhou community finished a series of questionnaire as the other experiment finished in Wenzhou.

***Procedure.***

The procedure was similar to Experiment 1. There were 6 blocks of trial, each with

<sub>1252</sub> 120 trials for 2017 data. Due to procedure error, the data collected in 2015 in Tsinghua  
<sub>1253</sub> community only have 60 trials for each block, i.e., 30 trials per condition.

<sub>1254</sub> As in study 3a, before each task, the instruction showed the meaning of each label to  
<sub>1255</sub> participants. The self-matching task and other-matching task were randomized between  
<sub>1256</sub> participants. Each participant finished 6 blocks, each have 120 trials.

<sub>1257</sub> ***Data Analysis.***

<sub>1258</sub> Same as experiment 3a.

<sub>1259</sub> **Results.**

<sub>1260</sub> ***NHST.***

<sub>1261</sub> Figure 27 shows  $d$  prime and reaction times of experiment 3a. Less than 5% correct  
<sub>1262</sub> trials with less than 200ms reaction times were excluded.

<sub>1263</sub>  $d$  prime.

<sub>1264</sub> There was no evidence for the main effect of valence,  $F(1.59, 69.94) = 2.34$ ,  
<sub>1265</sub>  $MSE = 0.48$ ,  $p = .115$ ,  $\hat{\eta}_G^2 = .010$ , but we found a main effect of self-relevance,  
<sub>1266</sub>  $F(1, 44) = 0.00$ ,  $MSE = 0.08$ ,  $p = .994$ ,  $\hat{\eta}_G^2 = .000$ , as well as the interaction,  
<sub>1267</sub>  $F(1.96, 86.41) = 0.53$ ,  $MSE = 0.10$ ,  $p = .585$ ,  $\hat{\eta}_G^2 = .001$ .

<sub>1268</sub> We then conducted separated ANOVA for self-referential and other-referential trials.  
<sub>1269</sub> The valence effect was shown for the self-referential conditions,  $F(1.75, 76.86) = 3.08$ ,  
<sub>1270</sub>  $MSE = 0.25$ ,  $p = .058$ ,  $\hat{\eta}_G^2 = .017$ . Post-hoc test revealed that the Good-Self condition  
<sub>1271</sub> ( $2.15 \pm 0.12$ ) was with greater  $d$  prime than Neutral condition ( $1.83 \pm 0.12$ ,  $t(34) = 3.36$ ,  
<sub>1272</sub>  $p = 0.0031$ ), and Bad-self condition ( $1.87 \pm 0.12$ ),  $t(34) = 2.955$ ,  $p = 0.01$ . There was  
<sub>1273</sub> difference between neutral and bad condition,  $t(34) = -0.039$ ,  $p = 0.914$ . However, no effect  
<sub>1274</sub> of valence was found for the other-referential condition  $F(1.63, 71.50) = 1.07$ ,  $MSE = 0.33$ ,  
<sub>1275</sub>  $p = .336$ ,  $\hat{\eta}_G^2 = .006$ .

<sub>1276</sub> *Reaction time.*

1277 We found interaction between Matchness and Valence ( $F(1.87, 82.50) = 18.58$ ,  
 1278  $MSE = 1,291.12$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .023$ ) and then analyzed the matched trials and  
 1279 nonmatch trials separately, as in previous experiments.

1280 For the match trials, we found that the interaction between identity and valence,  
 1281  $F(1.86, 81.84) = 5.22$ ,  $MSE = 308.30$ ,  $p = .009$ ,  $\hat{\eta}_G^2 = .003$ , as well as the main effect of  
 1282 valence  $F(1.80, 79.37) = 11.04$ ,  $MSE = 2,937.54$ ,  $p < .001$ ,  $\hat{\eta}_G^2 = .059$ , but not the effect of  
 1283 identity  $F(1, 44) = 0.23$ ,  $MSE = 263.26$ ,  $p = .632$ ,  $\hat{\eta}_G^2 = .000$ . As for the  $d$  prime, we  
 1284 separated analyzed the self-referential and other-referential trials. For the Self-referential  
 1285 trials, we found the main effect of valence,  $F(1.74, 76.48) = 13.69$ ,  $MSE = 1,732.08$ ,  
 1286  $p < .001$ ,  $\hat{\eta}_G^2 = .079$ ; for the other-referential trials, the effect of valence is weaker,  
 1287  $F(1.87, 82.44) = 7.09$ ,  $MSE = 1,527.43$ ,  $p = .002$ ,  $\hat{\eta}_G^2 = .043$ . We then focused on the self  
 1288 conditions: the good-self condition ( $713 \pm 12$ ) is faster than neutral- ( $776 \pm 11.8$ ),  $t(34) =$   
 1289  $-7.396$ ,  $p < .0001$ , and bad-self ( $772 \pm 10.1$ ) conditions,  $t(34) = -5.66$ ,  $p < .0001$ . But  
 1290 there is not difference between neutral- and bad-self conditions,  $t(34) = 0.481$ ,  $p = 0.881$ .

1291 For the nonmatch trials, we didn't found any strong effect: identity,  $F(1, 44) = 1.96$ ,  
 1292  $MSE = 319.47$ ,  $p = .169$ ,  $\hat{\eta}_G^2 = .001$ , valence  $F(1.69, 74.54) = 6.59$ ,  $MSE = 886.19$ ,  
 1293  $p = .004$ ,  $\hat{\eta}_G^2 = .010$ , or interaction between the two  $F(1.88, 82.57) = 0.31$ ,  $MSE = 316.96$ ,  
 1294  $p = .718$ ,  $\hat{\eta}_G^2 = .000$ .

1295 **BGLM.**

1296 *Signal detection theory analysis of accuracy.*

1297 We found that the  $d$  prime is greater when shapes were associated with good self  
 1298 condition than with neutral self or bad self, but shapes associated with bad self and neutral  
 1299 self didn't show differences. comparing the self vs other under three condition revealed that  
 1300 shapes associated with good self is greater than with good other, but with a weak evidence.  
 1301 In contrast, for both neutral and bad valence condition, shapes associated with other had  
 1302 greater  $d$  prime than with self.

### Reaction time.

In reaction times, we found that same trends in the match trials as in the RT: while the shapes associated with good self was greater than with good other (log mean diff = -0.02858, 95%HPD[-0.070898, 0.0154]), the direction is reversed for neutral and negative condition. see Figure 28

**HDDM.** We fitted our data with HDDM, using the response-coding (also see Hu et al., 2020). We estimated separate drift rate ( $v$ ), non-decision time ( $T_0$ ), and boundary separation ( $a$ ) for each condition. We found that the shapes tagged with good person has higher drift rate and higher boundary separation than shapes tagged with both neutral and bad person. Also, the shapes tagged with neutral person has a higher drift rate than shapes tagged with bad person, but not for the boundary separation. Finally, we found that shapes tagged with bad person had longer non-decision time (see figure 29)).

## Results

1316 Effect of moral valence

In this part, we synthesized results from experiment 1a, 1b, 1c, 2, 5 and 6a. Data from 192 participants were included in these analyses. We found differences between positive and negative conditions on RT was Cohen's  $d = -0.58 \pm 0.06$ , 95% CI [-0.70 -0.47]; on  $d'$  was Cohen's  $d = 0.24 \pm 0.05$ , 95% CI [0.15 0.34]. The effect was also observed between positive and neutral condition, RT: Cohen's  $d = -0.44 \pm 0.10$ , 95% CI [-0.63 -0.25];  $d'$ : Cohen's  $d = 0.31 \pm 0.07$ , 95% CI [0.16 0.45]. And the difference between neutral and bad conditions are not significant, RT: Cohen's  $d = 0.15 \pm 0.07$ , 95% CI [0.00 0.30];  $d'$ : Cohen's  $d = 0.07 \pm 0.07$ , 95% CI [-0.08 0.21]. See Figure 30 left panel.

**1325 Interaction between valence and self-reference**

1326 In this part, we combined the experiments that explicitly manipulated the  
1327 self-reference and valence, which includes 3a, 3b, 6b, 7a, and 7b. For the positive versus  
1328 negative contrast, data were from five experiments with 178 participants; for positive  
1329 versus neutral and neutral versus negative contrasts, data were from three experiments (1330 3a, 3b, and 6b) with 108 participants.

1331 In most of these experiments, the interaction between self-reference and valence was  
1332 significant (see results of each experiment in supplementary materials). In the  
1333 mini-meta-analysis, we analyzed the valence effect for self-referential condition and  
1334 other-referential condition separately.

1335 For the self-referential condition, we found the same pattern as in the first part of  
1336 results. That is we found significant differences between positive and neutral as well as  
1337 positive and negative, but not neutral and negative. The effect size of RT between positive  
1338 and negative is Cohen's  $d = -0.89 \pm 0.12$ , 95% CI [-1.11 -0.66]; on  $d'$  was Cohen's  $d = 0.61$   
1339  $\pm 0.09$ , 95% CI [0.44 0.78]. The effect was also observed between positive and neutral  
1340 condition, RT: Cohen's  $d = -0.76 \pm 0.13$ , 95% CI [-1.01 -0.50];  $d'$ : Cohen's  $d = 0.69 \pm$   
1341  $0.14$ , 95% CI [0.42 0.96]. And the difference between neutral and bad conditions are not  
1342 significant, RT: Cohen's  $d = 0.03 \pm 0.13$ , 95% CI [-0.22 0.29];  $d'$ : Cohen's  $d = 0.08 \pm 0.08$ ,  
1343 95% CI [-0.07 0.24]. See Figure 30 the middle panel.

1344 For the other-referential condition, we found that only the difference between positive  
1345 and negative on RT was significant, all the other conditions were not. The effect size of RT  
1346 between positive and negative is Cohen's  $d = -0.28 \pm 0.05$ , 95% CI [-0.38 -0.17]; on  $d'$  was  
1347 Cohen's  $d = -0.02 \pm 0.08$ , 95% CI [-0.17 0.13]. The effect was not observed between  
1348 positive and neutral condition, RT: Cohen's  $d = -0.12 \pm 0.10$ , 95% CI [-0.31 0.06];  $d'$ :  
1349 Cohen's  $d = 0.01 \pm 0.08$ , 95% CI [-0.16 0.17]. And the difference between neutral and bad  
1350 conditions are not significant, RT: Cohen's  $d = 0.14 \pm 0.09$ , 95% CI [-0.03 0.31];  $d'$ :

<sub>1351</sub> Cohen's  $d = 0.05 \pm 0.07$ , 95% CI [-0.08 0.18]. See Figure 30 right panel.

<sub>1352</sub> **Generalizability of the valence effect**

<sub>1353</sub> In this part, we reported the results from experiment 4 in which either moral valence  
<sub>1354</sub> or self-reference were manipulated as task-irrelevant stimuli.

<sub>1355</sub> For experiment 4a, when self-reference was the target and moral valence was  
<sub>1356</sub> task-irrelevant, we found that only under the implicit self-referential condition, i.e., when  
<sub>1357</sub> the moral words were presented as task irrelevant stimuli, there was the main effect of  
<sub>1358</sub> valence and interaction between valence and reference for both  $d$  prime and RT (See  
<sub>1359</sub> supplementary results for the detailed statistics). For  $d$  prime, we found good-self  
<sub>1360</sub> condition ( $2.55 \pm 0.86$ ) had higher  $d$  prime than bad-self condition ( $2.38 \pm 0.80$ ); good self  
<sub>1361</sub> condition was also higher than neutral self ( $2.45 \pm 0.78$ ) but there was not statistically  
<sub>1362</sub> significant, while the neutral-self condition was higher than bad self condition and not  
<sub>1363</sub> significant neither. For reaction times, good-self condition ( $654.26 \pm 67.09$ ) were faster  
<sub>1364</sub> relative to bad-self condition ( $665.64 \pm 64.59$ ), and over neutral-self condition ( $664.26 \pm$   
<sub>1365</sub>  $64.71$ ). The difference between neutral-self and bad-self conditions were not significant.  
<sub>1366</sub> However, for the other-referential condition, there was no significant differences between  
<sub>1367</sub> different valence conditions. See Figure 31.

<sub>1368</sub> For experiment 4b, when valence was the target and the identity was task-irrelevant,  
<sub>1369</sub> we found a strong valence effect (see supplementary results and Figure 32, Figure 33).

<sub>1370</sub> In this experiment, the advantage of good-self condition can only be disentangled by  
<sub>1371</sub> comparing the self-referential and other-referential conditions. Therefore, we calculated the  
<sub>1372</sub> differences between the valence effect under self-referential and other referential conditions  
<sub>1373</sub> and used the weighted variance as the variance of this differences. We found this  
<sub>1374</sub> modulation effect on RT. The valence effect of RT was stronger in self-referential than  
<sub>1375</sub> other-referential for the Good vs. Neutral condition ( $-0.33 \pm 0.01$ ), and to a less extent the

<sup>1376</sup> Good vs. Bad condition ( $-0.17 \pm 0.01$ ). While the size of the other effect's CI included  
<sup>1377</sup> zero, suggesting those effects didn't differ from zero. See Figure 34.

<sup>1378</sup> **Specificity of valence effect**

<sup>1379</sup> In this part, we analyzed the results from experiment 5, which included positive,  
<sup>1380</sup> neutral, and negative valence from four different domains: morality, emotion, aesthetics of  
<sup>1381</sup> human, and aesthetics of scene. We found interaction between valence and domain for both  
<sup>1382</sup> *d* prime and RT (match trials). A common pattern appeared in all four domains: each  
<sup>1383</sup> domain showed a binary results instead of gradient on both *d* prime and RT. For morality,  
<sup>1384</sup> aesthetics of human, and aesthetics of scene, the positive conditions had advantages over  
<sup>1385</sup> both neutral and negative conditions (greater *d* prime and faster RT), and neutral and  
<sup>1386</sup> negative conditions didn't differ from each other. But for the emotional stimuli, it was the  
<sup>1387</sup> positive and neutral had advantage over negative conditions, while positive and neutral  
<sup>1388</sup> conditions were not significantly different. See supplementary materials for detailed  
<sup>1389</sup> statistics. Also note that the effect size in moral domain is smaller than the aesthetic  
<sup>1390</sup> domains (beauty of people and beauty of scene). See Figure 35.

<sup>1391</sup> **Self-reported personal distance**

<sup>1392</sup> See Figure 36.

<sup>1393</sup> **Correlation analyses**

<sup>1394</sup> The reliability of questionnaires can be found in (Liu et al., 2020). We calculated the  
<sup>1395</sup> correlation between the data from behavioral task and the questionnaire data.

<sup>1396</sup> We focused on the task-questionnaire correlation, the results revealed that the score  
<sup>1397</sup> from three questionnaire are related to behavioral responses data. First, the external moral  
<sup>1398</sup> identity is positively correlated with boundary separation of moral good condition,

<sub>1399</sub>  $r = 0.194$ , 95% CI [0.023 0.350]); the moral self image is positively correlated with the drift  
<sub>1400</sub> rate ( $r = 0.191$ , 95% CI [-0.016 0.354]) of the morally good condition. See Figure 37.

<sub>1401</sub> Second, we found the personal distance between self and good is positively correlated  
<sub>1402</sub> with the boundary separation of neutral condition and the self-neutral distance is  
<sub>1403</sub> negatively correlated with the boundary separation of neutral condition. See figure 38

<sub>1404</sub> Third, we found the self esteem score was negative correlated with the  $d'$  of bad  
<sub>1405</sub> conditions ( $r = -0.16$ , 95% CI [-0.277 -0.038]) and the neutral conditions ( $r = -.197$ , 95%  
<sub>1406</sub> CI [-0.348 -0.026]). See Figure 39.

<sub>1407</sub> We also explored the correlation between behavioral data and questionnaire scores  
<sub>1408</sub> separately for experiments with and without self-referential. For experiments without  
<sub>1409</sub> self-referential (Valence effect), we found the personal distance between Good-person and  
<sub>1410</sub> self is positively correlated with boundary separation of good conditions,  $r = 0.292$ , 95%  
<sub>1411</sub> [0.071 0.485]. also personal distance between the bad and neutral person is positively  
<sub>1412</sub> correlated with non-responding time of bad and neutral conditions,  $r = 0.249$ , 0.233,  
<sub>1413</sub> respectively.

<sub>1414</sub> For experiments with self-referential (Valence effect for the self), we found self-esteem  
<sub>1415</sub> is negatively correlated with  $d'$  prime of neutral condition,  $r = -0.272$ , [-0.468 -0.052], the  
<sub>1416</sub> self-good distance is positively correlated with  $d'$  prime for Bad condition,  $r = 0.185$ ,  
<sub>1417</sub> 95%CI[0.004 0.354].

## <sub>1418</sub> Discussion

## <sub>1419</sub> References

- <sub>1420</sub> Anderson, E., Siegel, E. H., Bliss-Moreau, E., & Barrett, L. F. (2011). The visual impact  
<sub>1421</sub> of gossip. *Science*, 332(6036), 1446–1448. <https://doi.org/10.1126/science.1201574>
- <sub>1422</sub> Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10(4), 433–436.

- 1423 Journal Article.
- 1424 Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan.  
1425 *Journal of Statistical Software; Vol 1, Issue 1 (2017)*. Journal Article. Retrieved  
1426 from  
1427 <https://www.jstatsoft.org/v080/i01>
- 1428 Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ...  
1429 Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of  
1430 Statistical Software, 76(1)*. Journal Article. <https://doi.org/10.18637/jss.v076.i01>
- 1431 Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis  
1432 and meta-analysis* (2nd ed.). Book, New York: Sage.
- 1433 DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological  
1434 Methods, 3(2)*, 186–205. Journal Article. <https://doi.org/10.1037/1082-989X.3.2.186>
- 1435 Farrell, B. (1985). "Same"—"different" judgments: A review of current controversies in  
1436 perceptual comparisons. *Psychological Bulletin, 98(3)*, 419–456. Journal Article.  
1437 <https://doi.org/10.1037/0033-2909.98.3.419>
- 1438 Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using  
1439 g\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research  
1440 Methods, 41(4)*, 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- 1441 Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced  
1442 perceptual awareness of morally relevant stimuli. *Cognition, 132(1)*, 22–29.  
1443 <https://doi.org/10.1016/j.cognition.2014.02.007>
- 1444 Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies:  
1445 Some arguments on why and a primer on how. *Social and Personality Psychology  
1446 Compass, 10(10)*, 535–549. Journal Article. <https://doi.org/10.1111/spc3.12267>
- 1447 Hu, C.-P., Lan, Y., Macrae, C. N., & Sui, J. (2020). Good me bad me: Does valence

- 1448 influence self-prioritization during perceptual decision-making? *Collabra: Psychology*, 6(1), 20. Journal Article. <https://doi.org/10.1525/collabra.301>
- 1449
- 1450 Kirby, K. N., & Gerlanc, D. (2013). BootES: An r package for bootstrap confidence  
1451 intervals on effect sizes. *Behavior Research Methods*, 45(4), 905–927.  
1452 <https://doi.org/10.3758/s13428-013-0330-5>
- 1453 Krueger, L. E. (1978). A theory of perceptual matching. *Psychological Review*, 85(4),  
1454 278–304. Journal Article. <https://doi.org/10.1037/0033-295X.85.4.278>
- 1455 Liu, Q., Wang, F., Yan, W., Peng, K., Sui, J., & Hu, C.-P. (2020). Questionnaire data from  
1456 the revision of a chinese version of free will and determinism plus scale. *Journal of  
1457 Open Psychology Data*, 8(1), 1. Journal Article. <https://doi.org/10.5334/jopd.49/>
- 1458 Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-gaussian  
1459 and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin &  
1460 Review*, 16(5), 798–817. <https://doi.org/10.3758/PBR.16.5.798>
- 1461 Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming  
1462 numbers into movies. *Spatial Vision*, 10(4), 437–442. Journal Article.
- 1463 Rouder, J. N., & Lu, J. (2005). An introduction to bayesian hierarchical models with an  
1464 application in the theory of signal detection. *Psychonomic Bulletin & Review*,  
1465 12(4), 573–604. Journal Article. <https://doi.org/10.3758/bf03196750>
- 1466 Rousselet, G. A., & Wilcox, R. R. (2019). Reaction times and other skewed distributions:  
1467 Problems with the mean and the median. *Meta-Psychology*. preprint.  
1468 <https://doi.org/10.1101/383935>
- 1469 Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. Conference  
1470 Proceedings. <https://doi.org/10.2139/ssrn.2205186>
- 1471 Spruyt, A., & Houwer, J. D. (2017). On the automaticity of relational stimulus processing:  
1472 The (extrinsic) relational simon task. *PLoS One*, 12(10), e0186606. Journal Article.

- 1473 <https://doi.org/10.1371/journal.pone.0186606>
- 1474 Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures.  
1475 *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. Journal  
1476 Article. <https://doi.org/10.3758/BF03207704>
- 1477 Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence  
1478 from self-prioritization effects on perceptual matching. *Journal of Experimental  
1479 Psychology: Human Perception and Performance*, 38(5), 1105–1117. Journal  
1480 Article. <https://doi.org/10.1037/a0029792>
- 1481 Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time  
1482 models applied to perceptual matching. *Psychonomic Bulletin & Review*, 7(2),  
1483 208–256. <https://doi.org/10.3758/BF03212980>
- 1484 Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of  
1485 the drift-diffusion model in python. *Frontiers in Neuroinformatics*, 7.  
1486 <https://doi.org/10.3389/fninf.2013.00014>

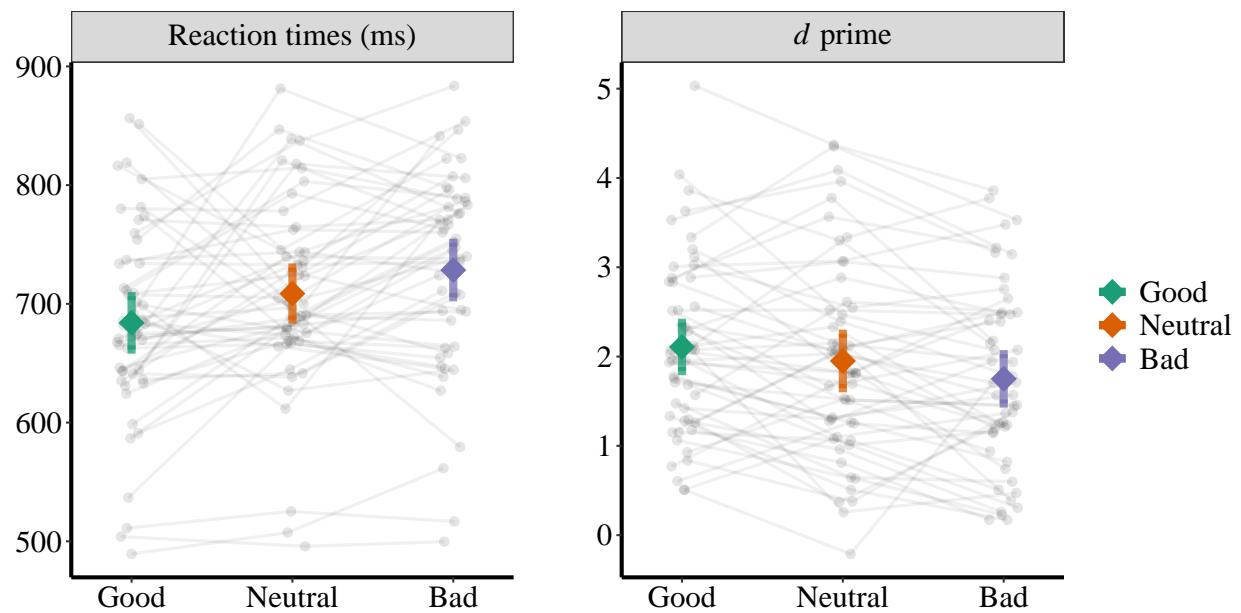


Figure 1. RT and  $d$  prime of Experiment 1a.

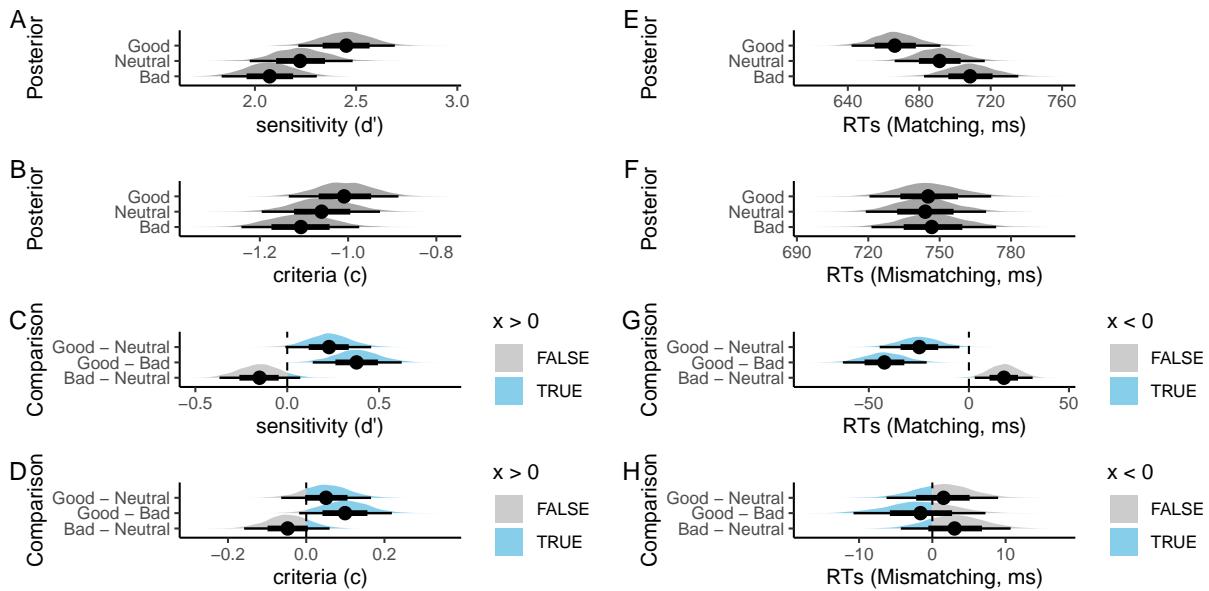


Figure 2. Exp1a: Results of Bayesian GLM analysis.

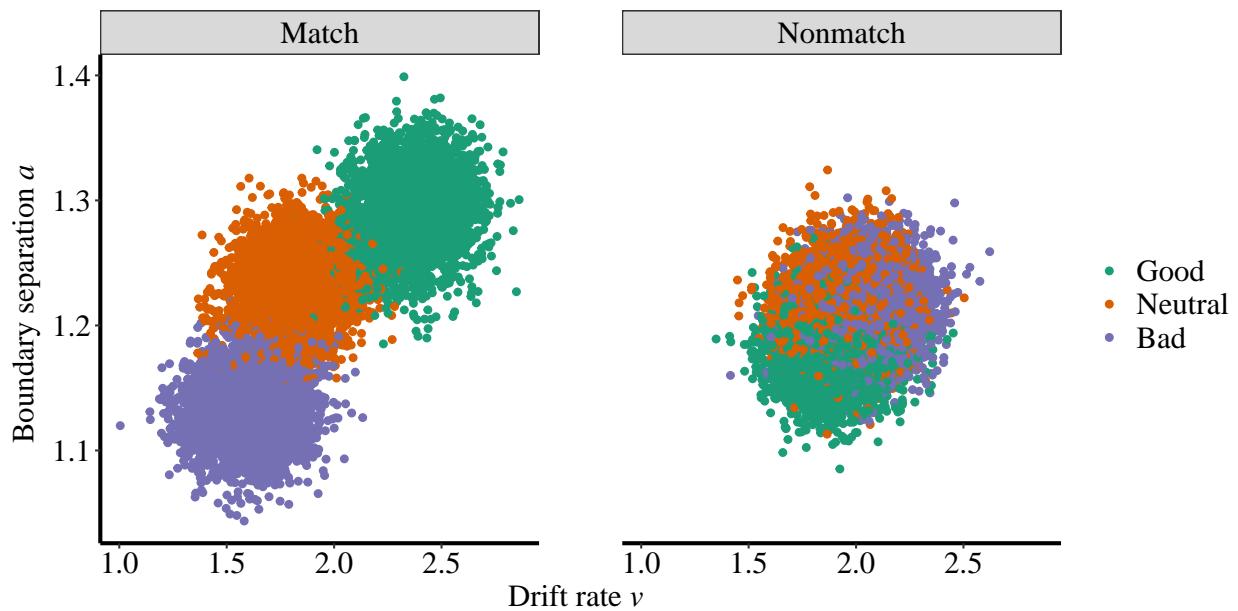


Figure 3. Exp1a: Results of HDDM.

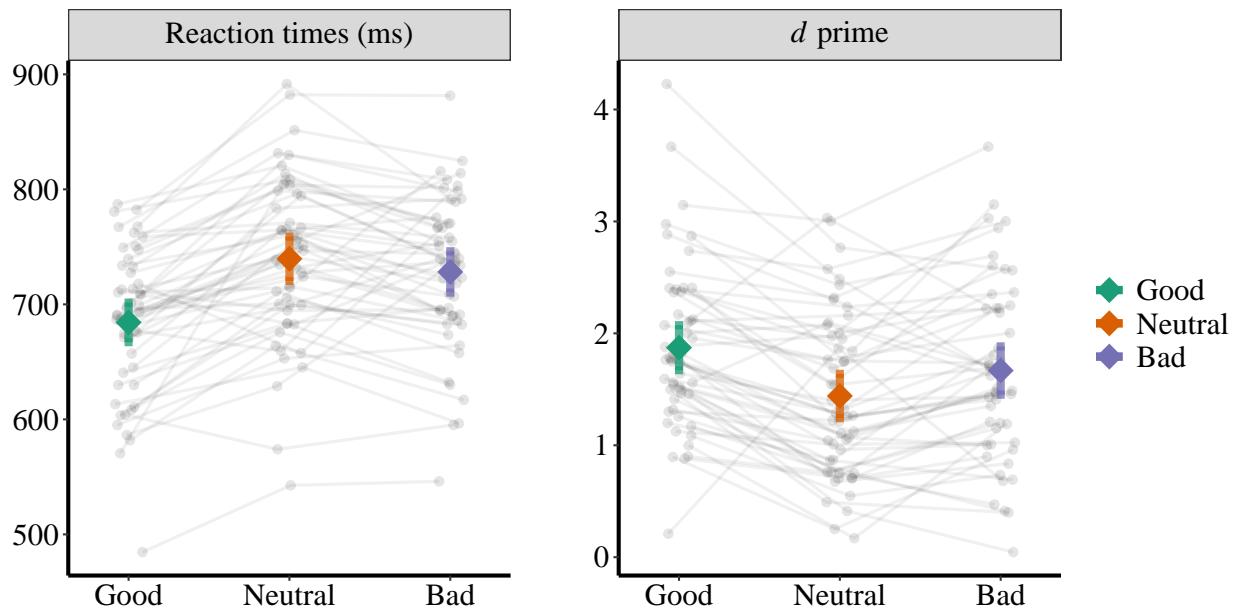


Figure 4. RT and *d* prime of Experiment 1b.

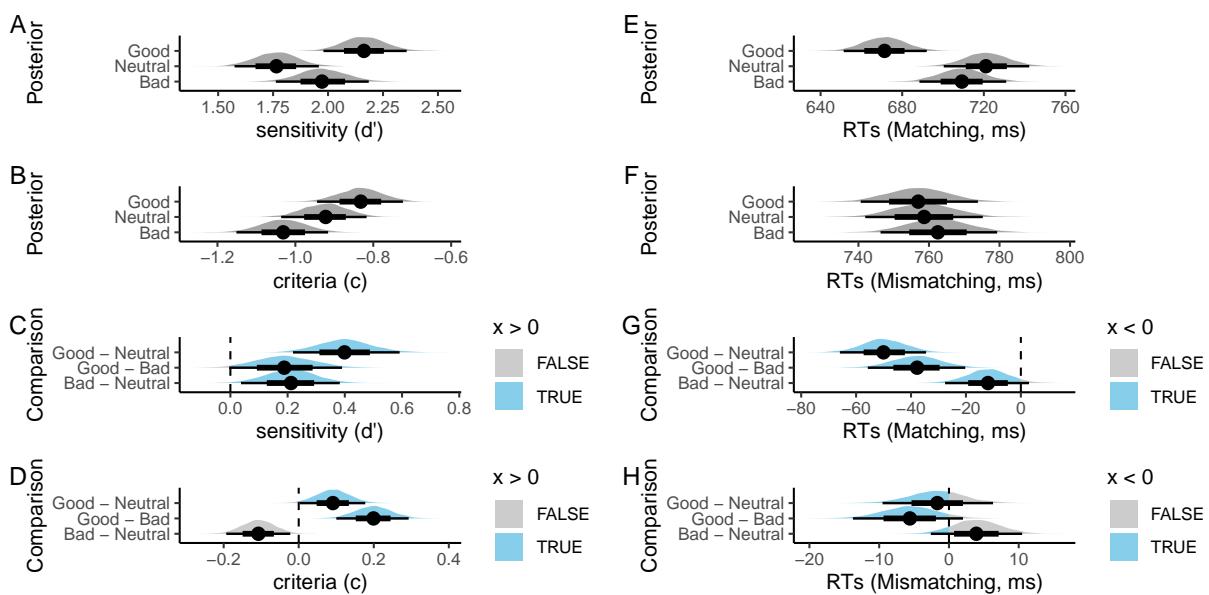


Figure 5. Exp1b: Results of Bayesian GLM analysis.

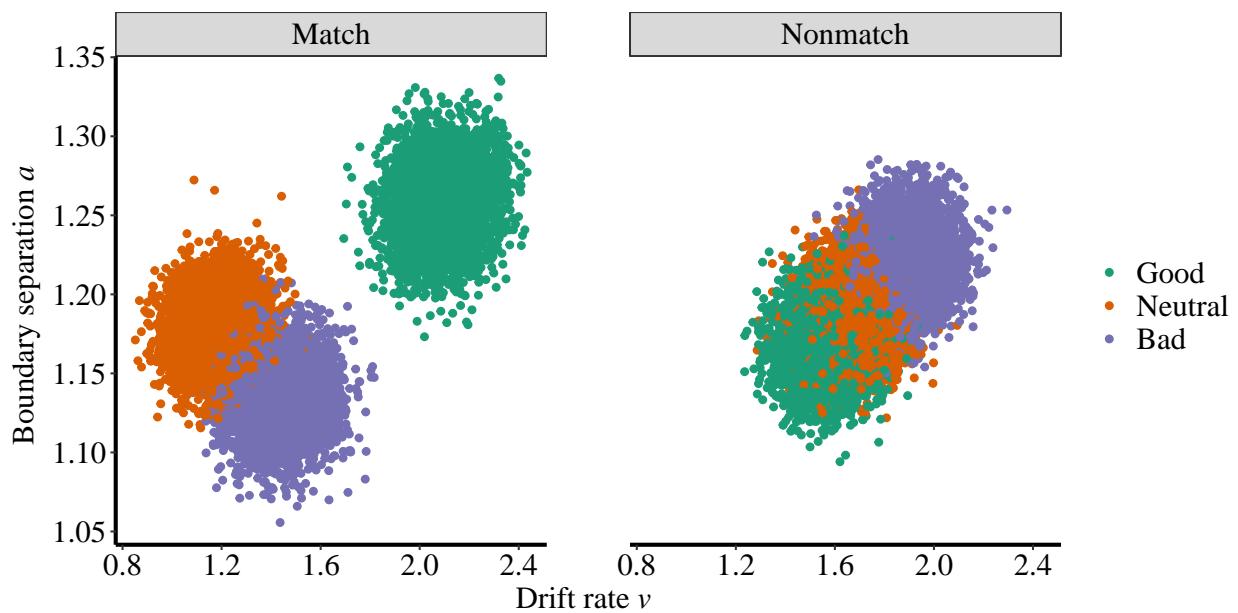


Figure 6. Exp1b: Results of HDDM.

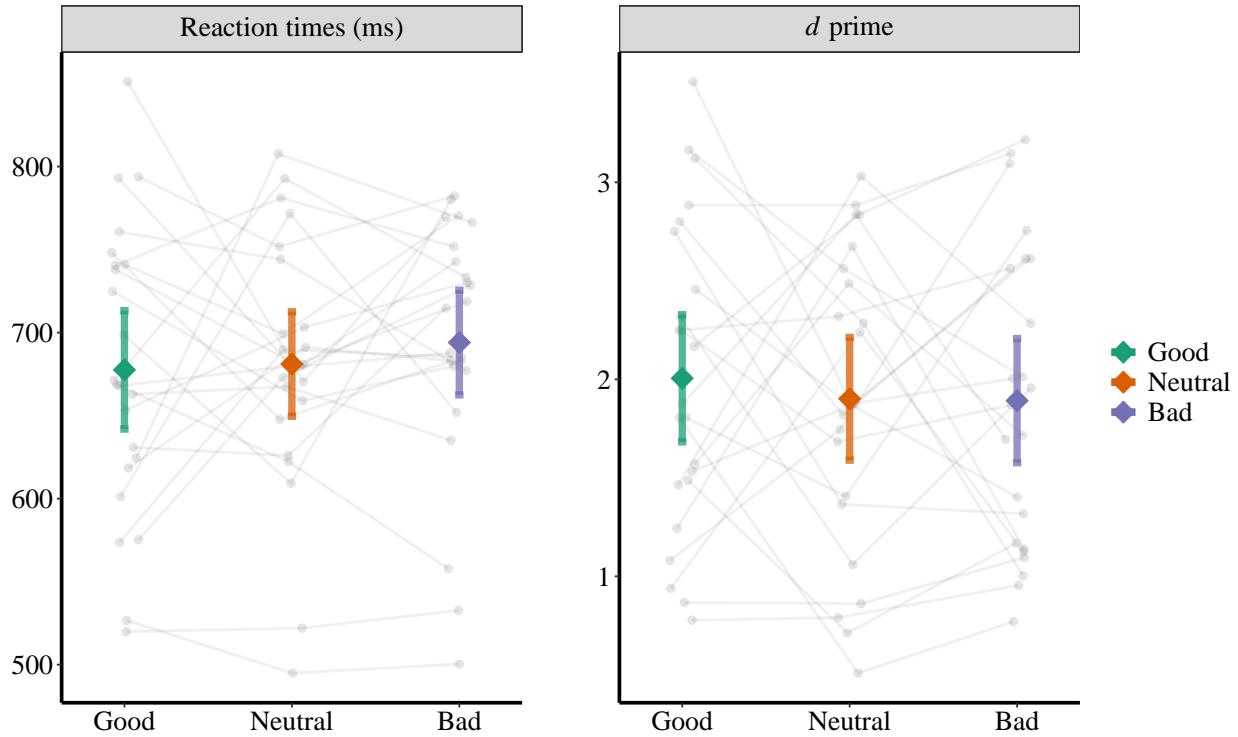


Figure 7. RT and  $d'$  prime of Experiment 1c.

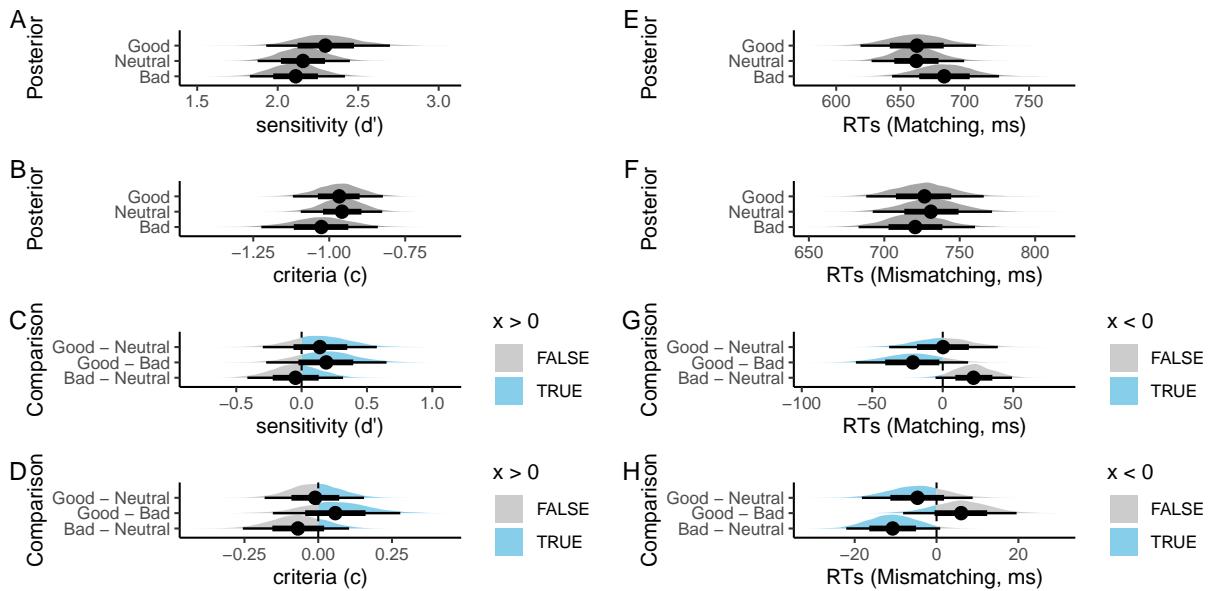


Figure 8. Exp1c: Results of Bayesian GLM analysis.

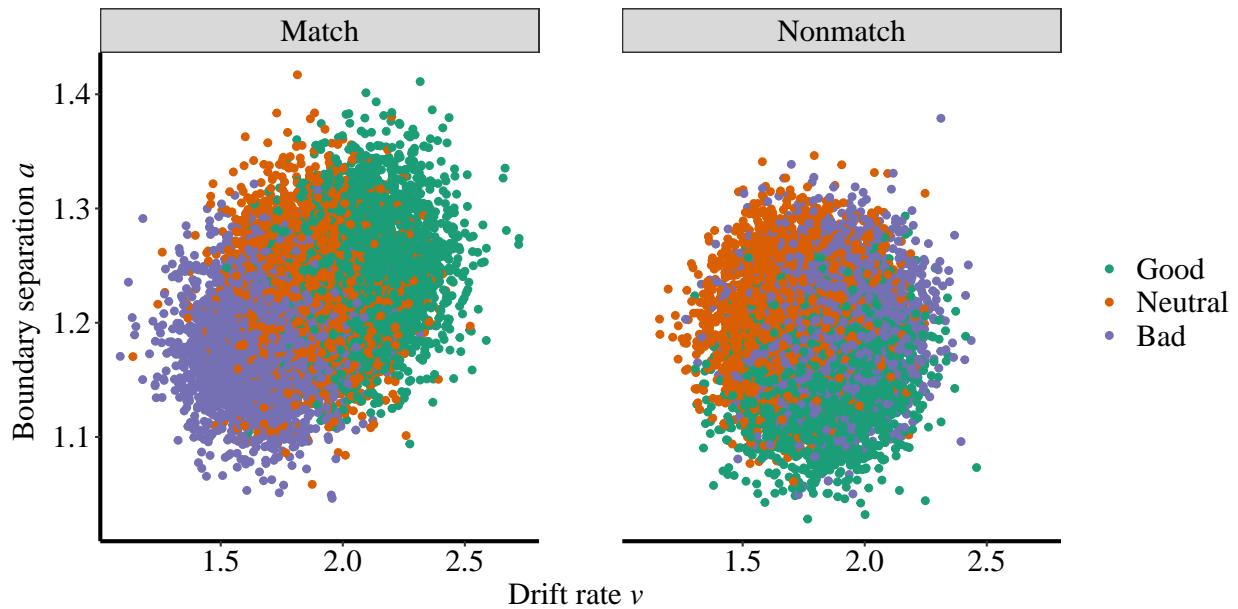


Figure 9. Exp1c: Results of HDDM.

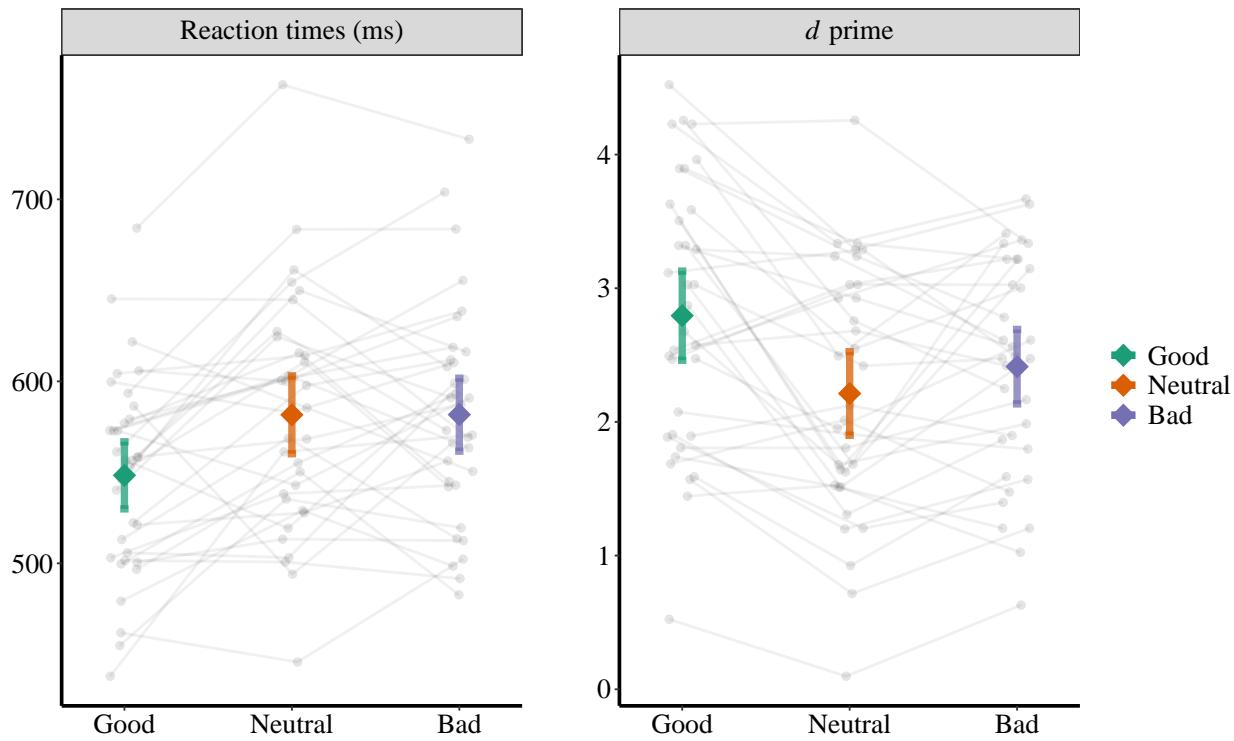


Figure 10. RT and  $d$  prime of Experiment 2.

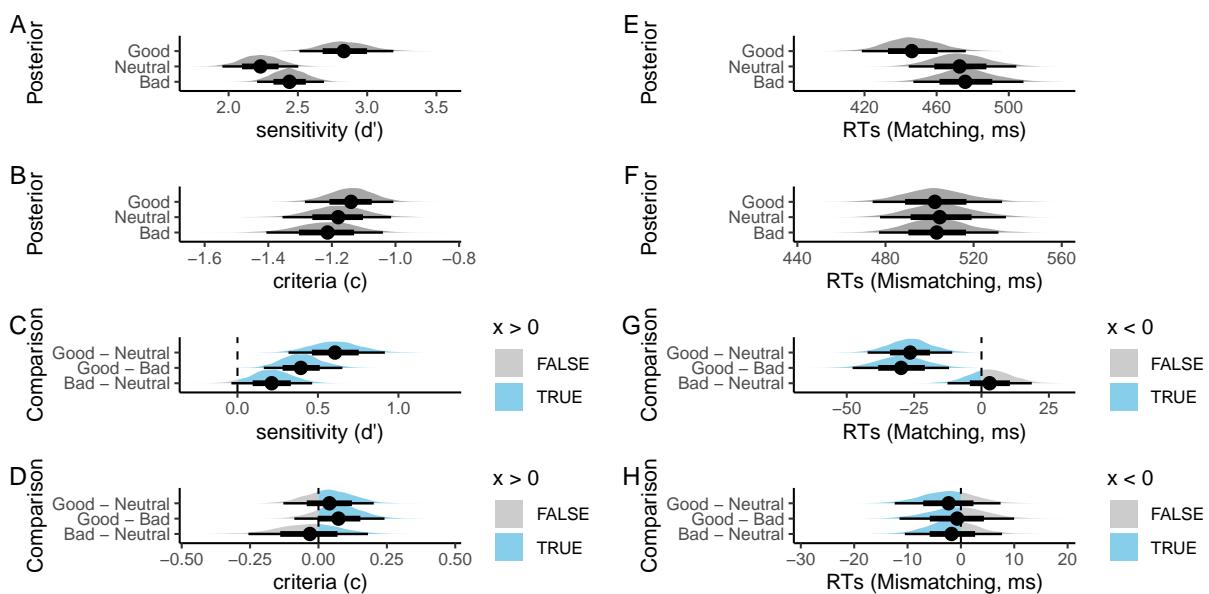


Figure 11. Exp2: Results of Bayesian GLM analysis.

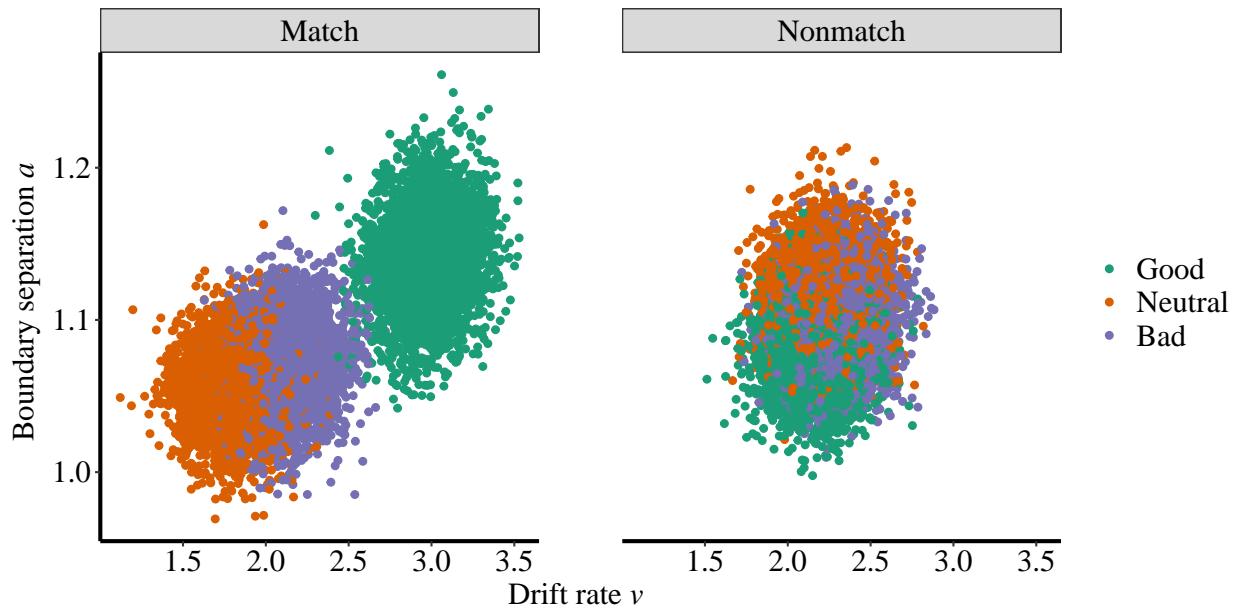


Figure 12. Exp2: Results of HDDM.

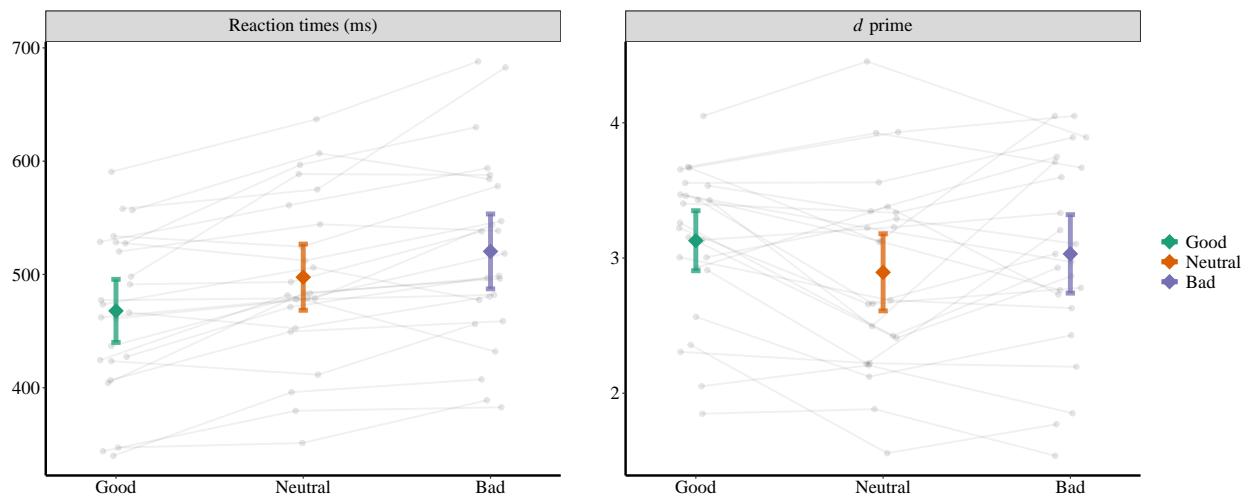


Figure 13. RT and  $d'$  prime of Experiment 6a.

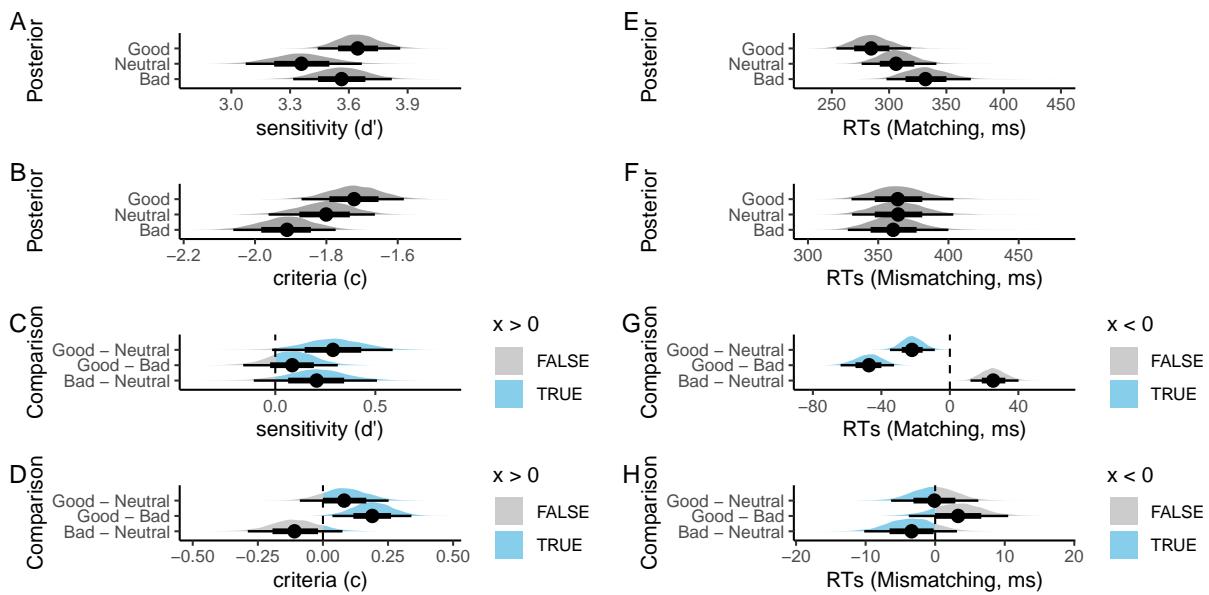


Figure 14. Exp1b: Results of Bayesian GLM analysis.

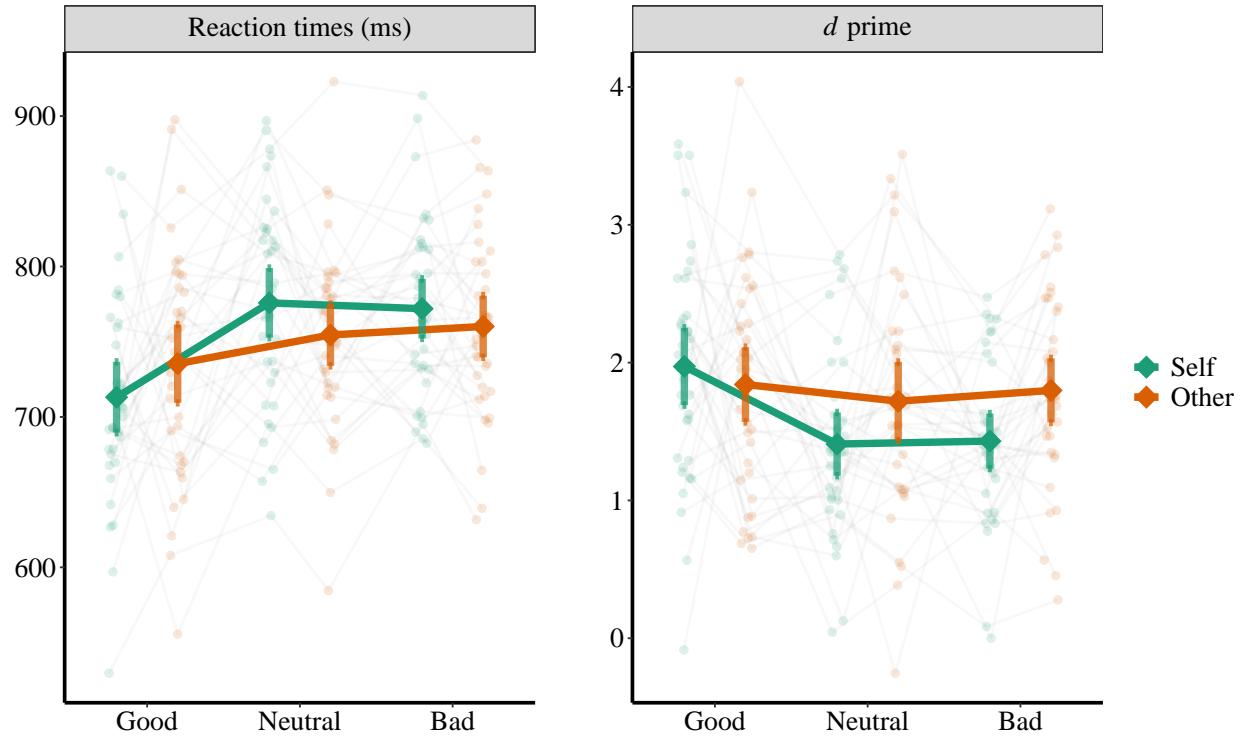


Figure 15. RT and  $d$  prime of Experiment 3a.

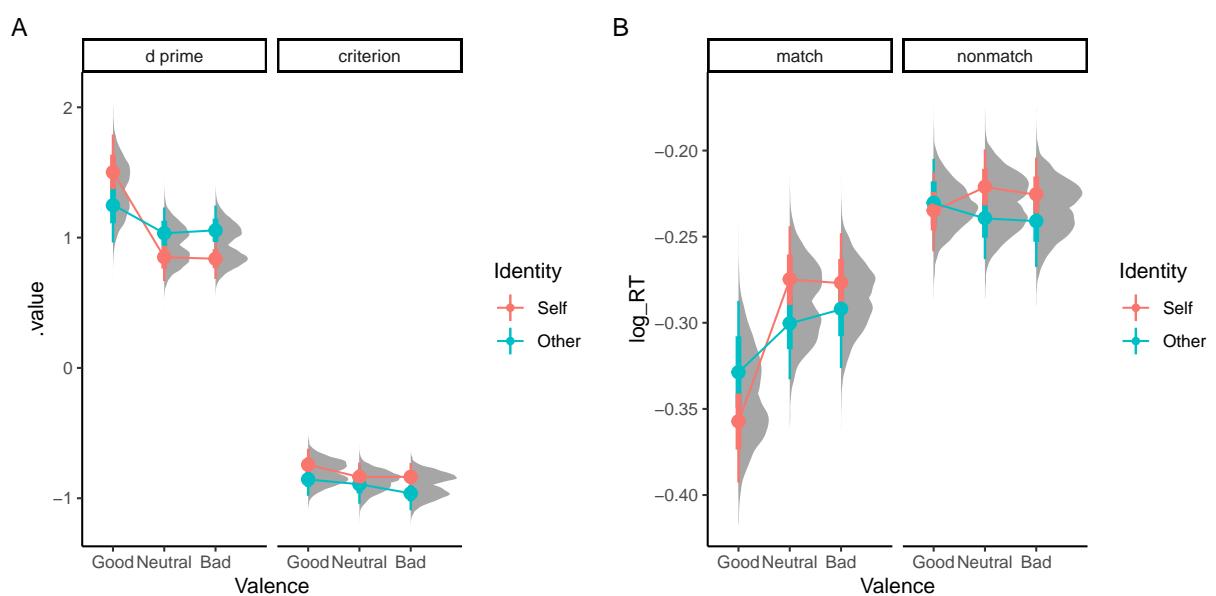


Figure 16. Exp3a: Results of Bayesian GLM analysis.

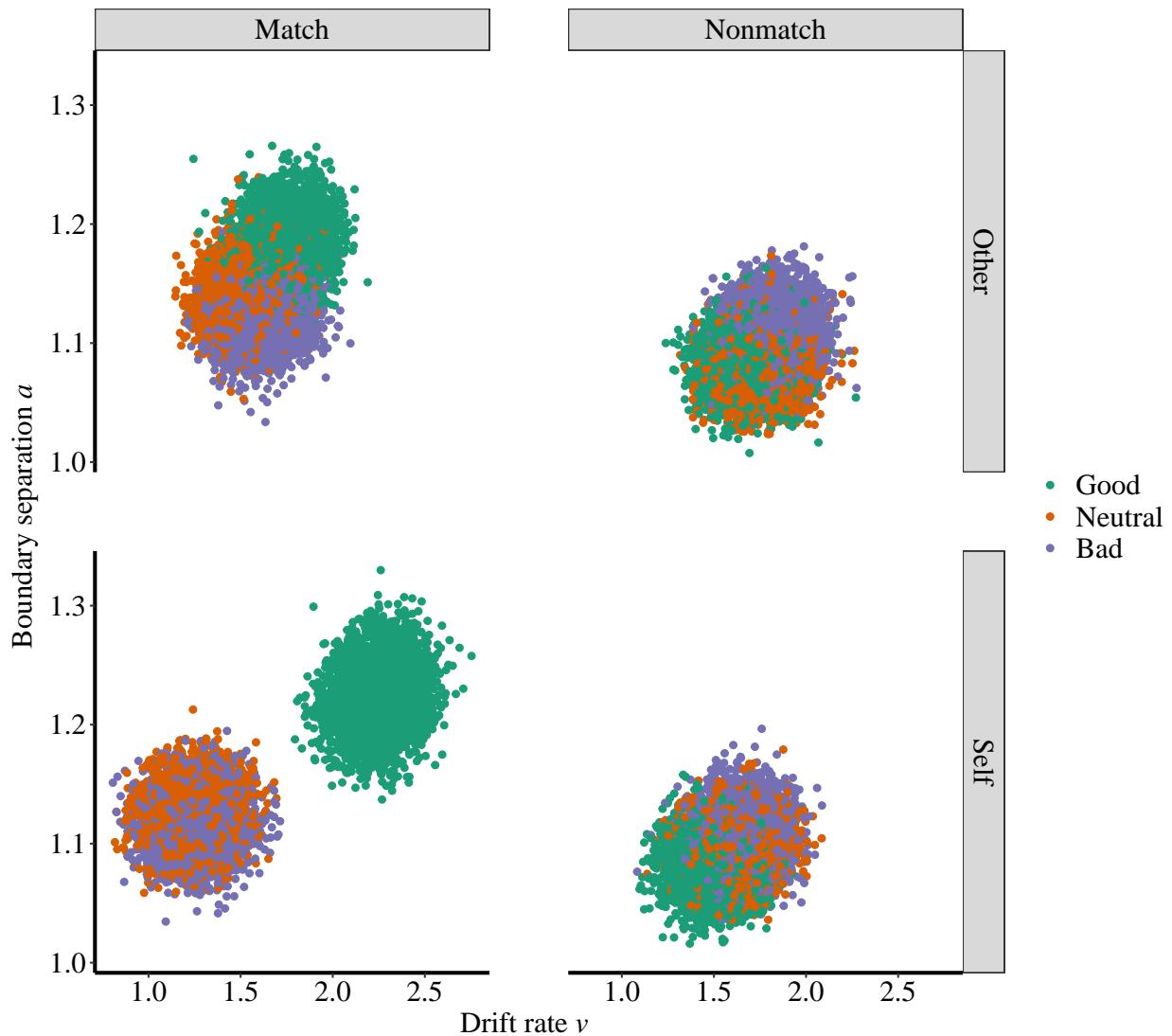


Figure 17. Exp3a: Results of HDDM.

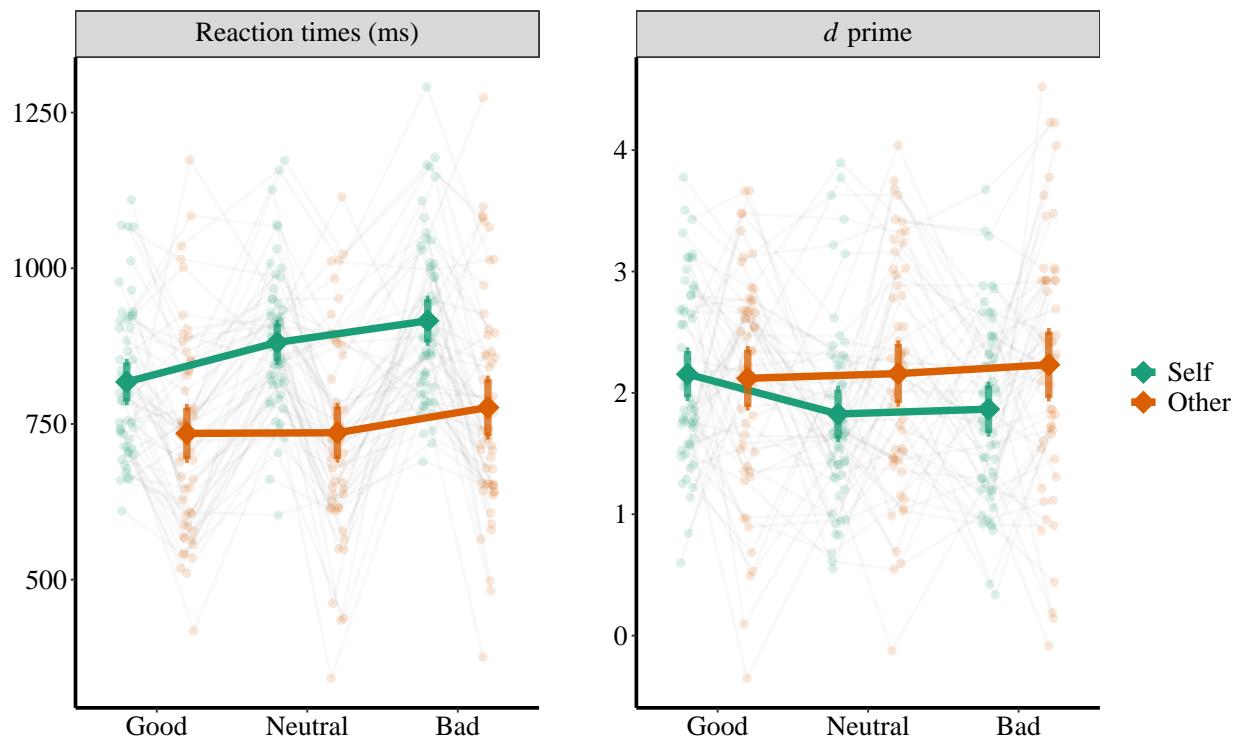


Figure 18. RT and  $d'$  of Experiment 3b.

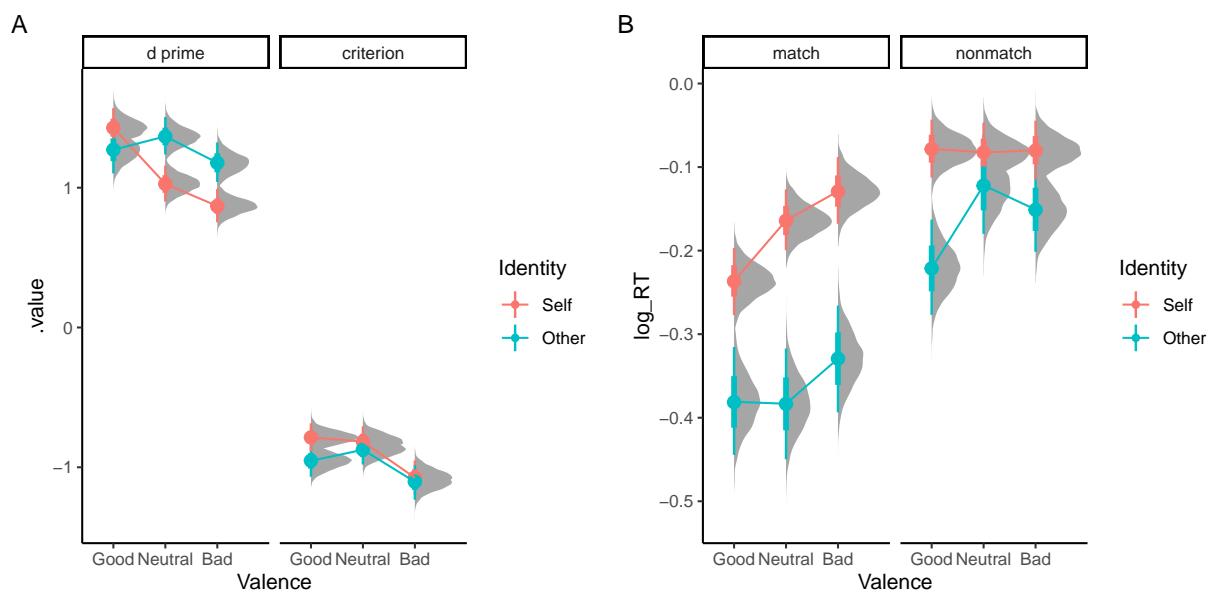


Figure 19. exp3b: Results of Bayesian GLM analysis.

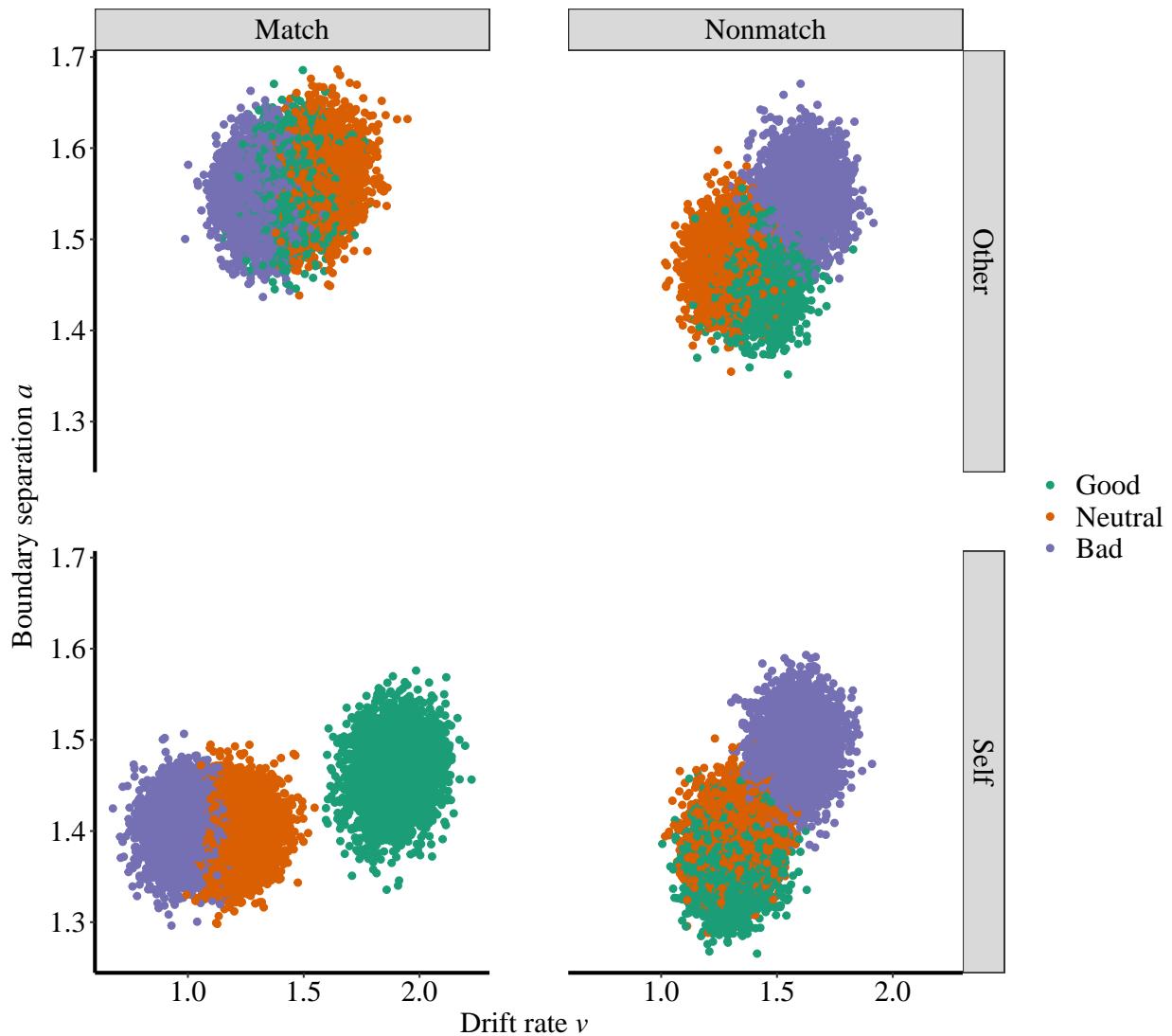


Figure 20. exp3b: Results of HDDM.

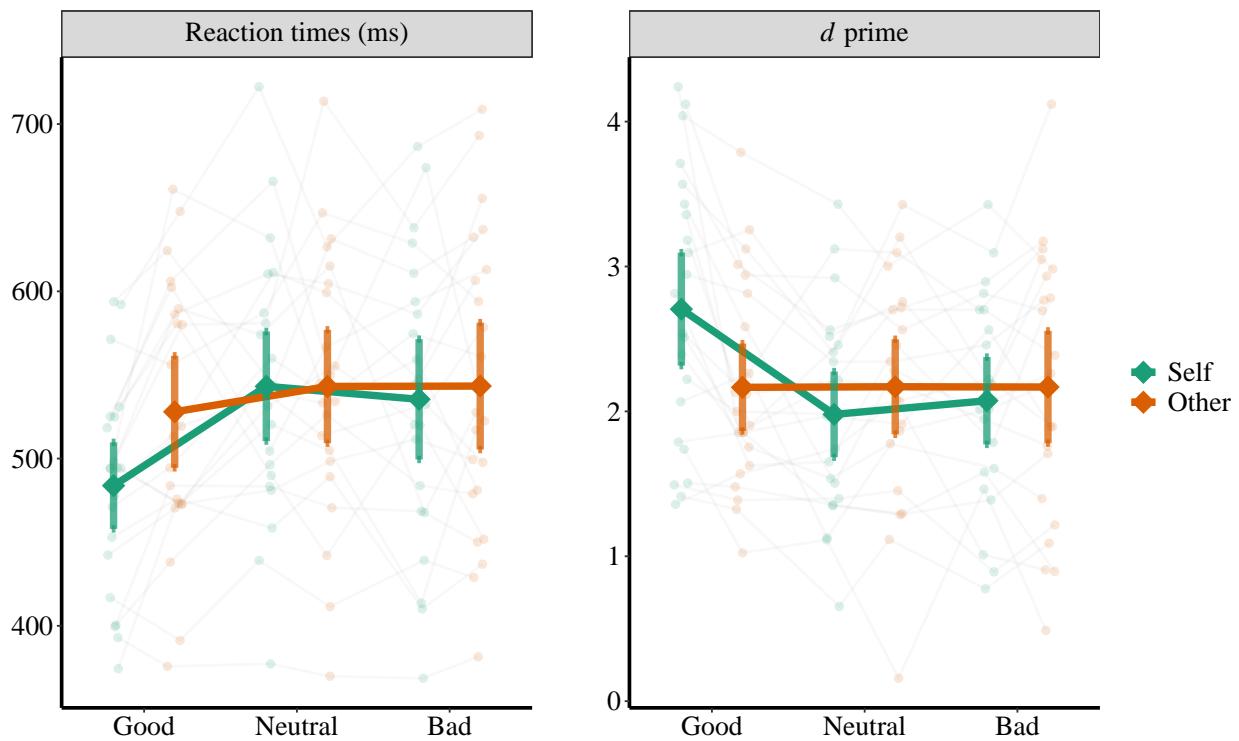
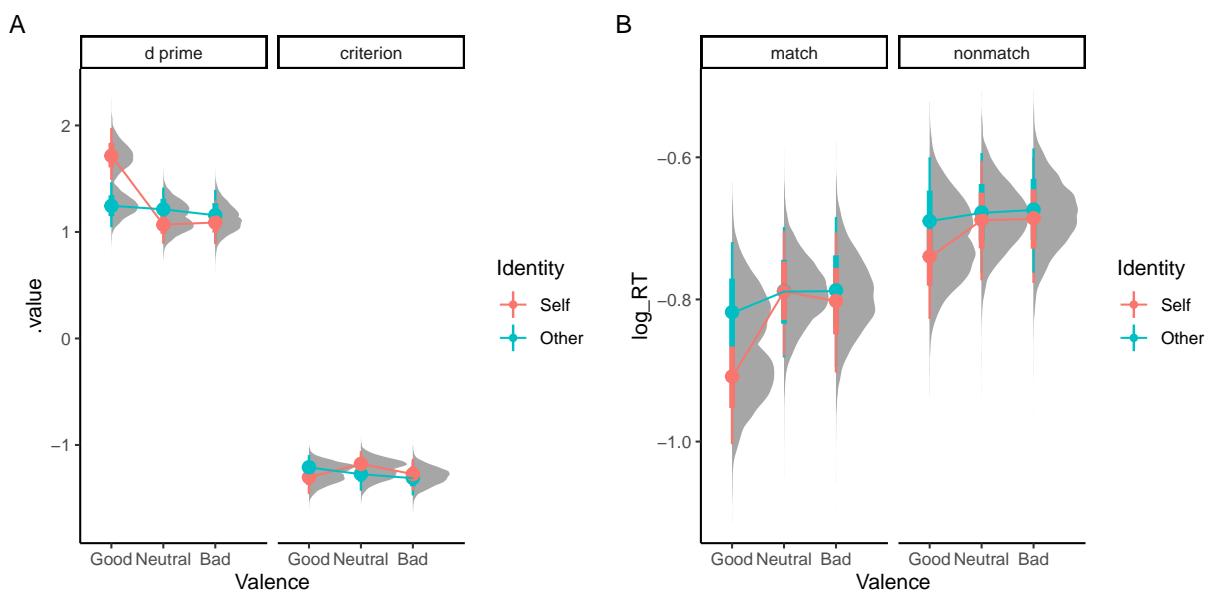
Figure 21. RT and  $d$  prime of Experiment 6b.

Figure 22. exp6b\_d1: Results of Bayesian GLM analysis.

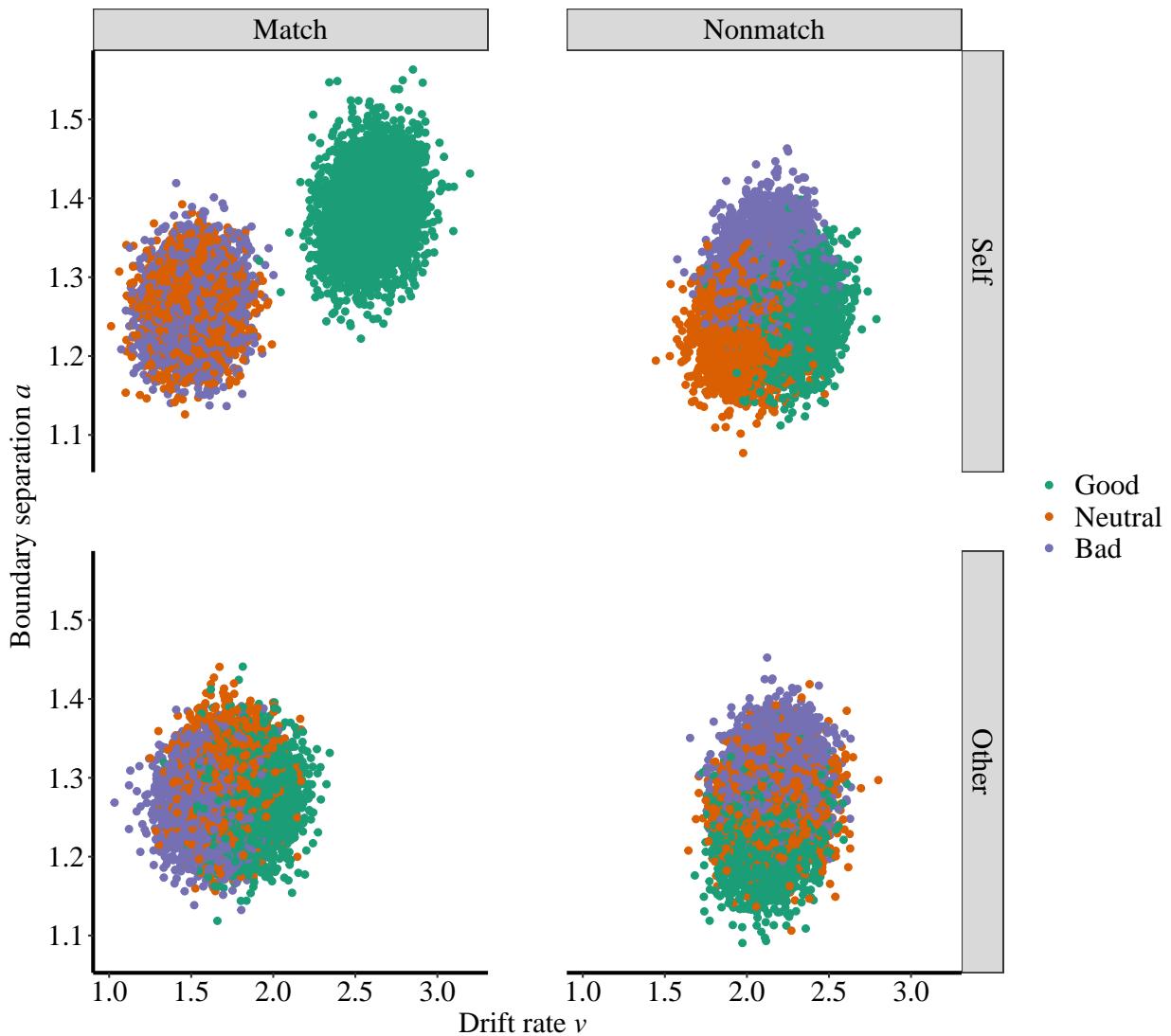


Figure 23. exp6b: Results of HDDM (Day 1).

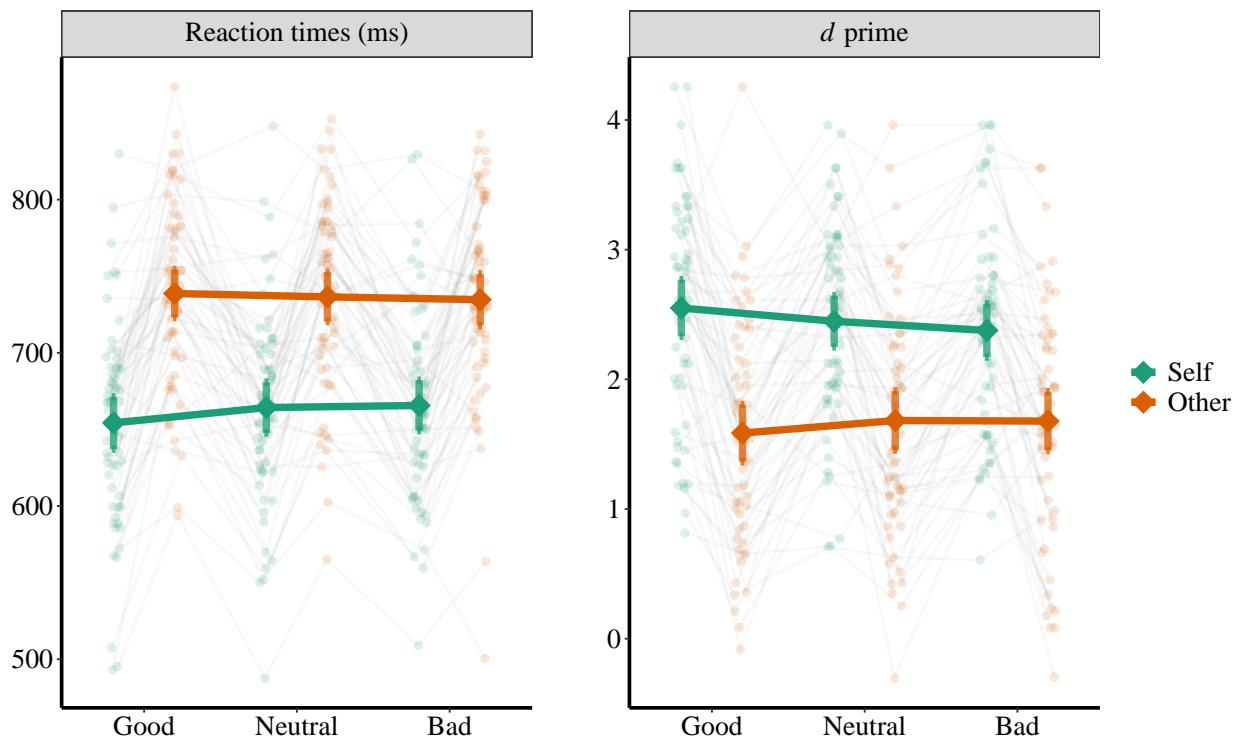


Figure 24. RT and  $d'$  of Experiment 4a.

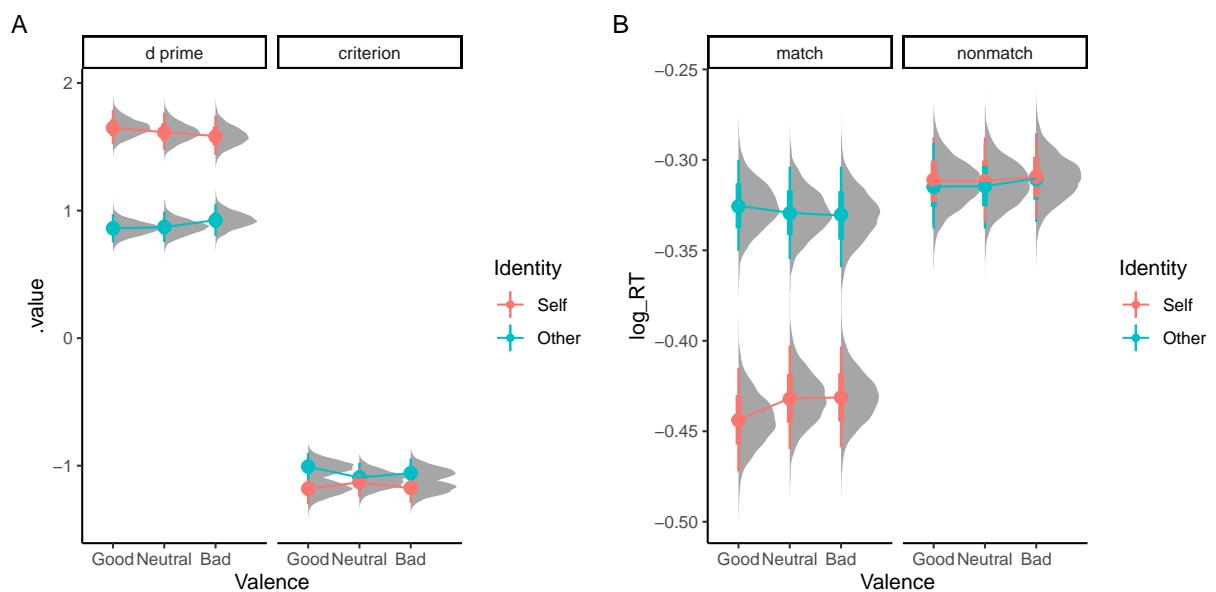


Figure 25. exp4a: Results of Bayesian GLM analysis.

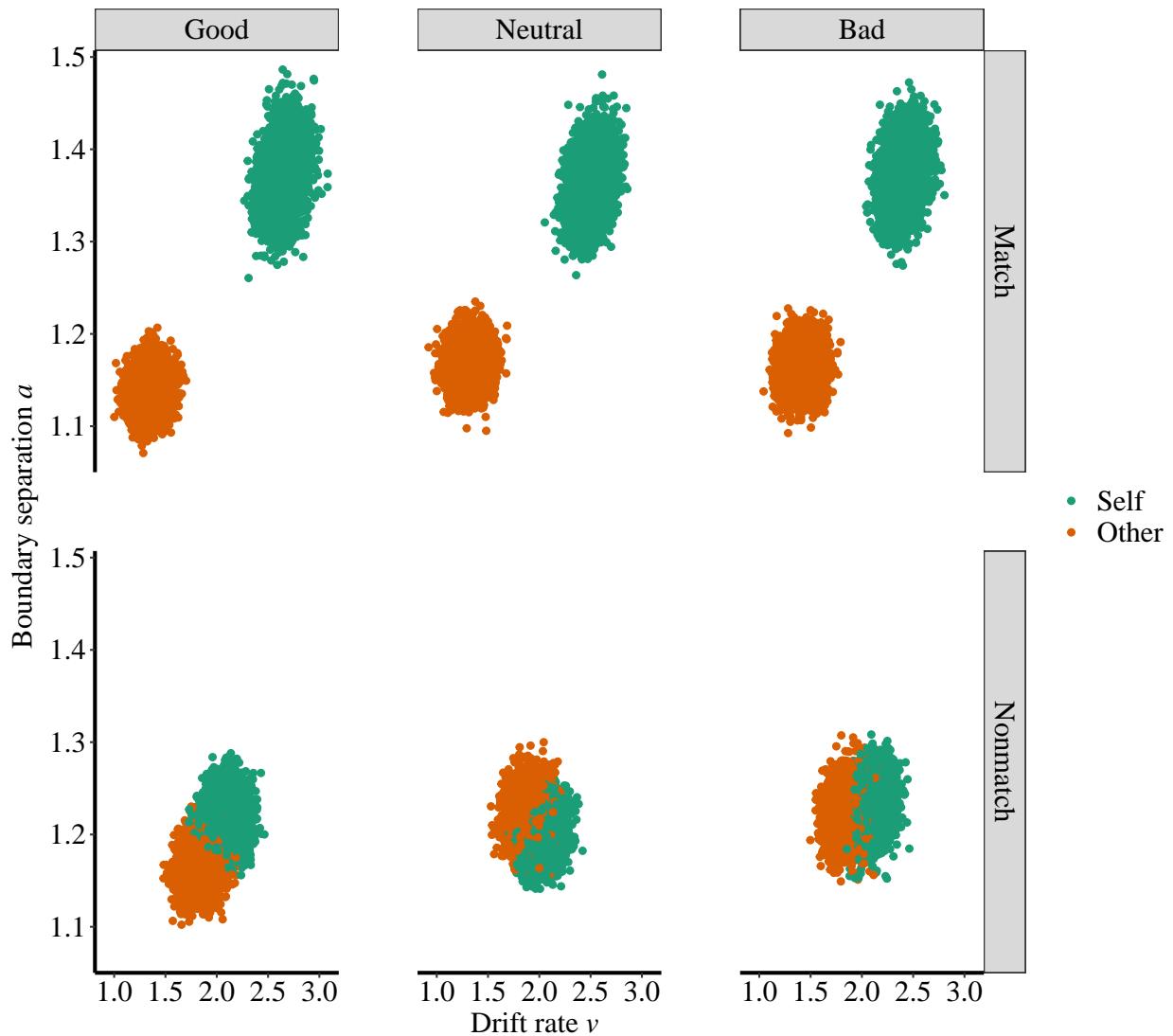


Figure 26. exp4a: Results of HDDM.

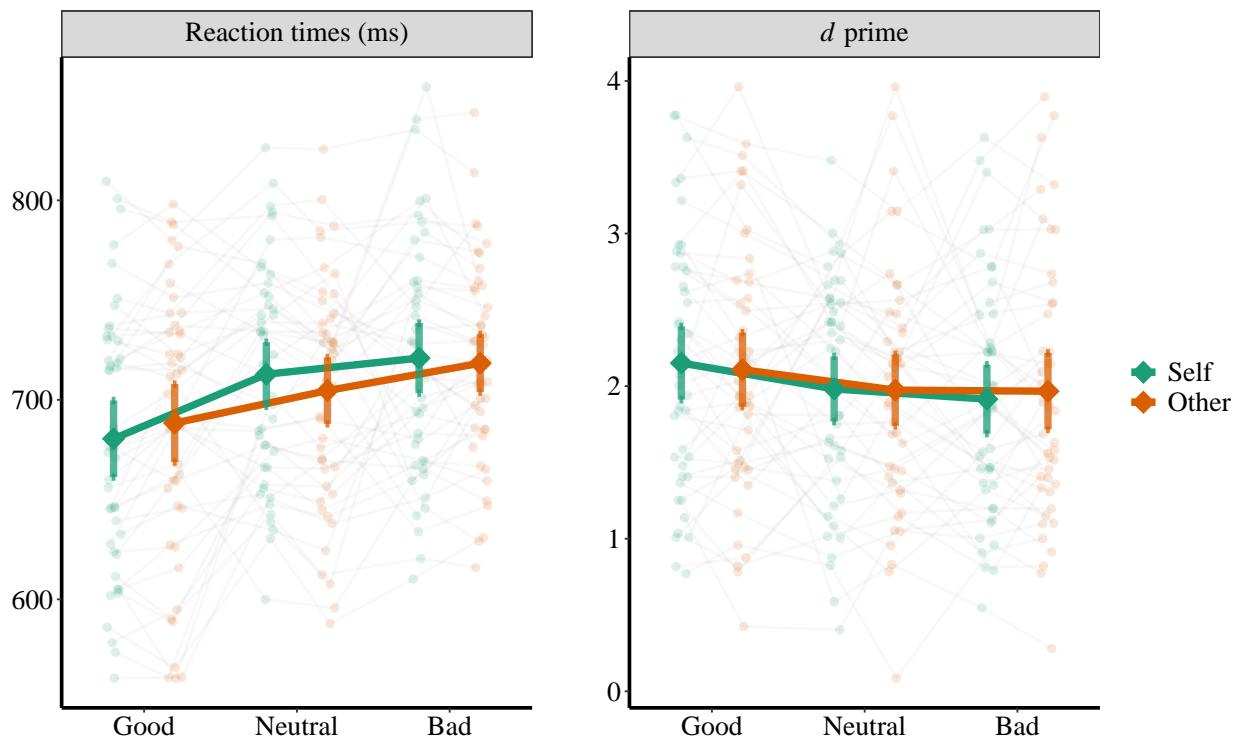


Figure 27. RT and  $d'$  of Experiment 4b.

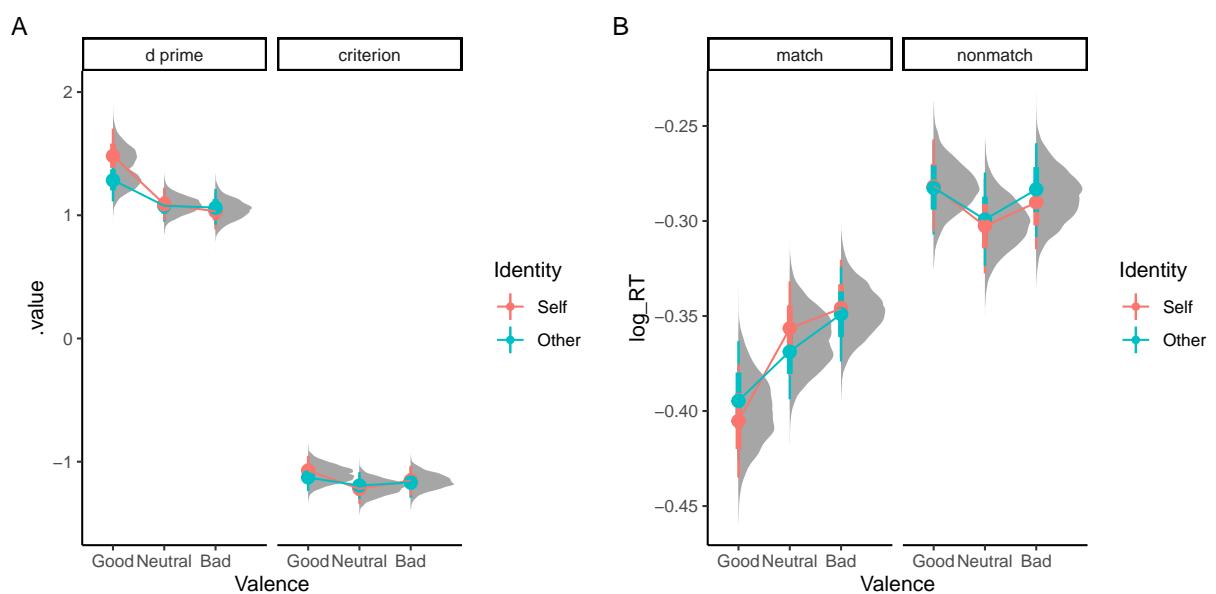


Figure 28. exp4b: Results of Bayesian GLM analysis.

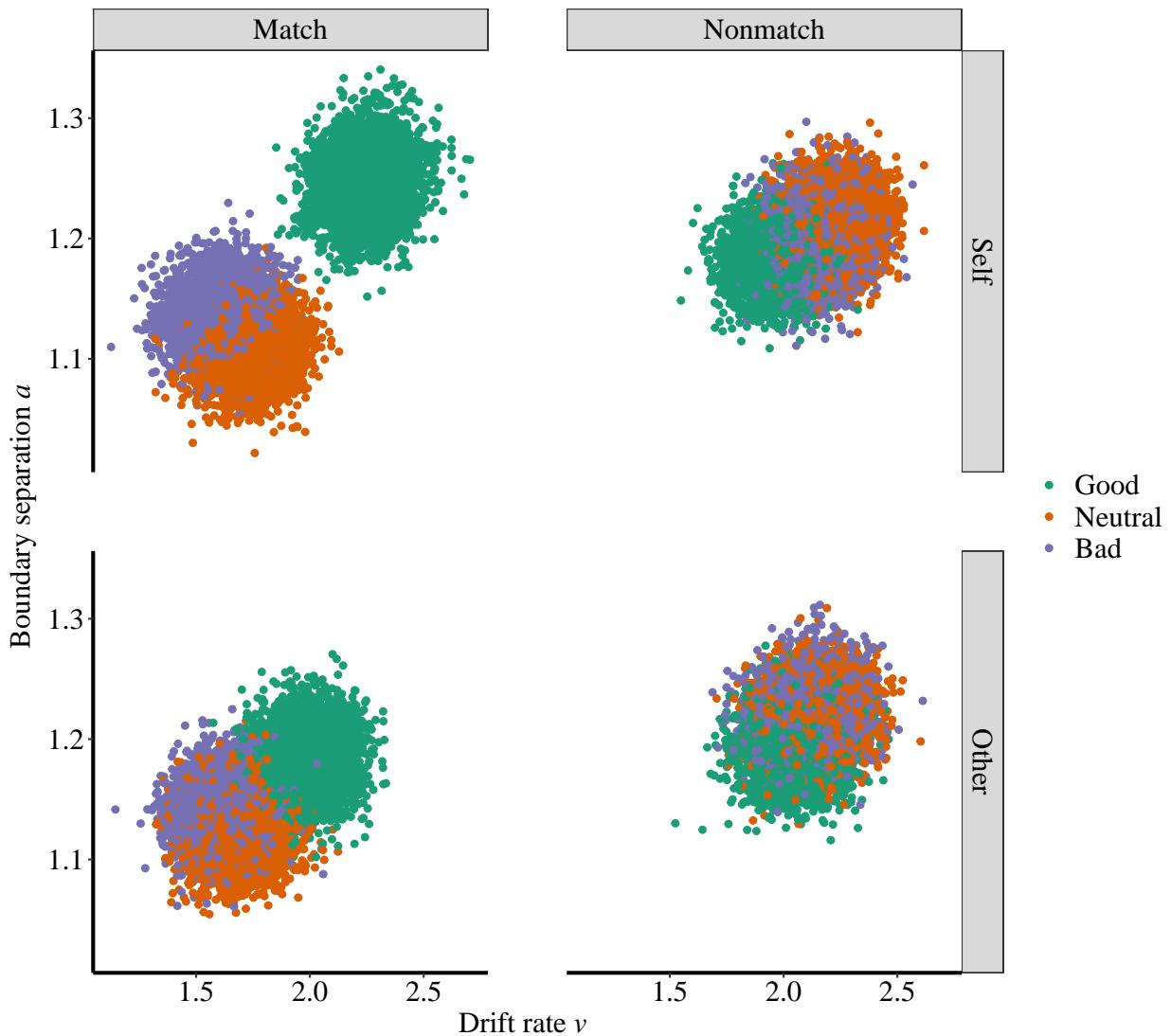


Figure 29. exp4b: Results of HDDM.

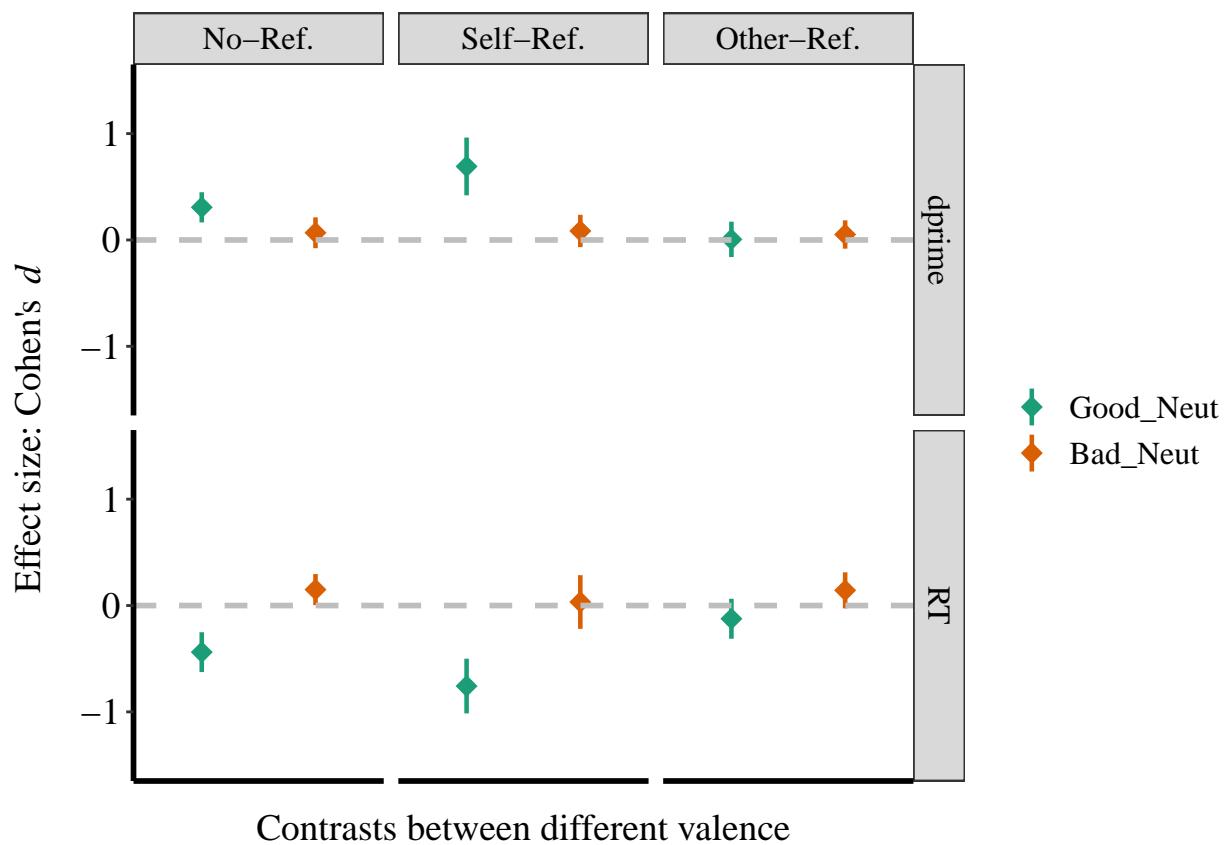


Figure 30. Effect size (Cohen's  $d$ ) of Valence.

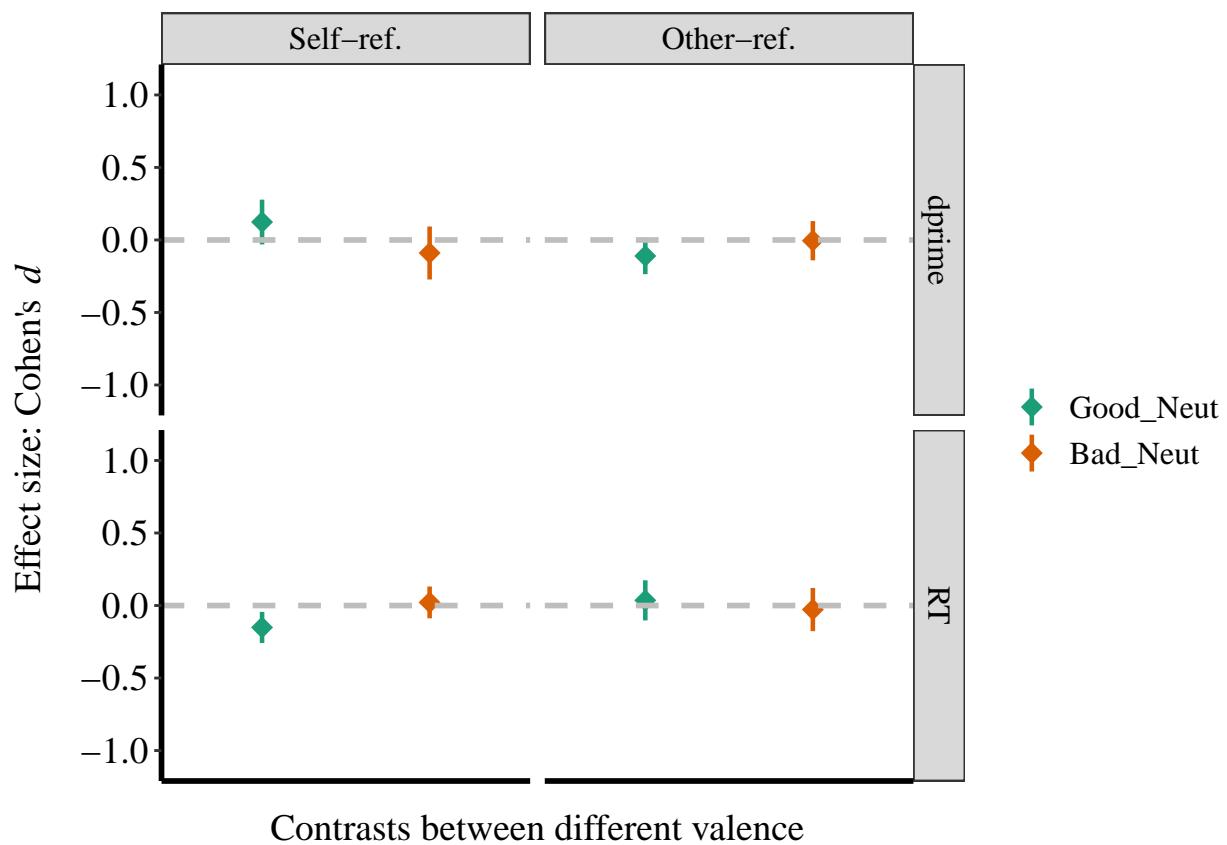


Figure 31. Effect size (Cohen's  $d$ ) of Valence in Exp4a.

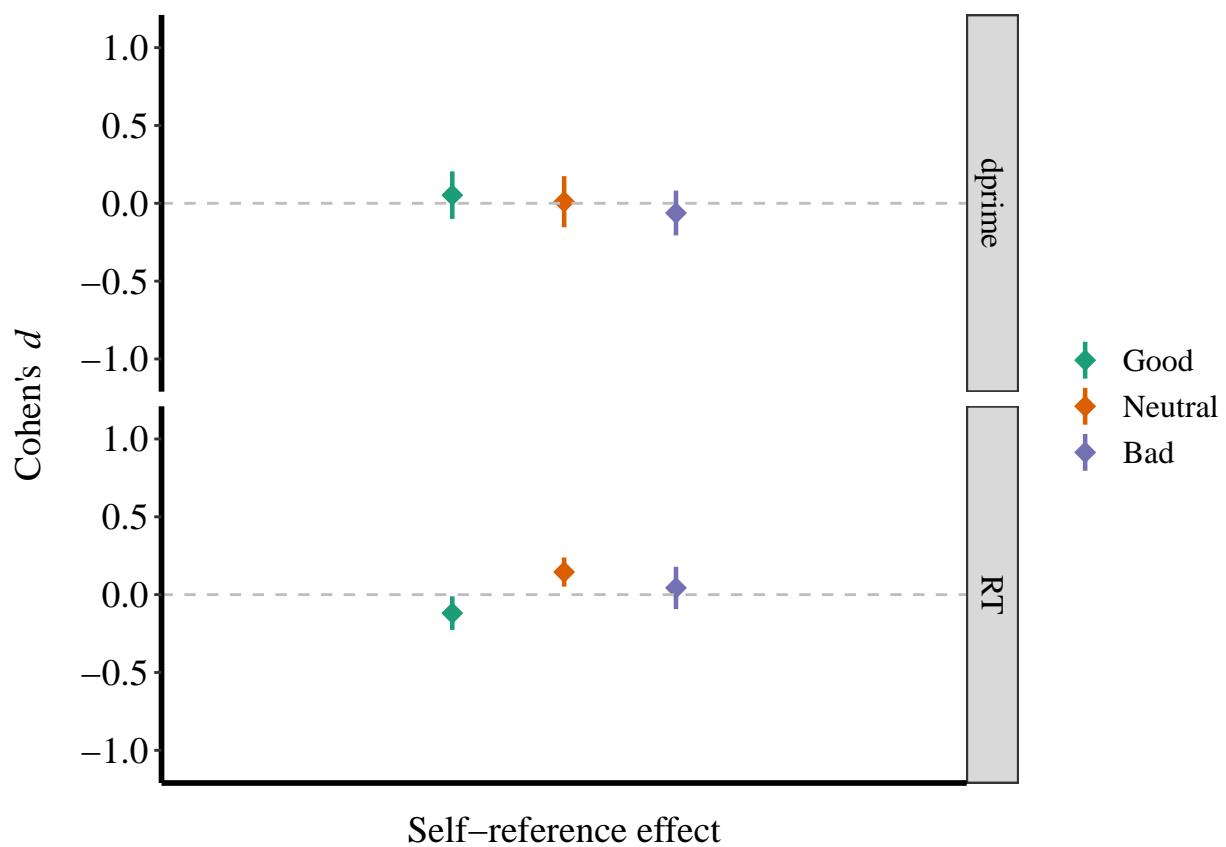


Figure 32. Effect size (Cohen's  $d$ ) of Valence in Exp4b.

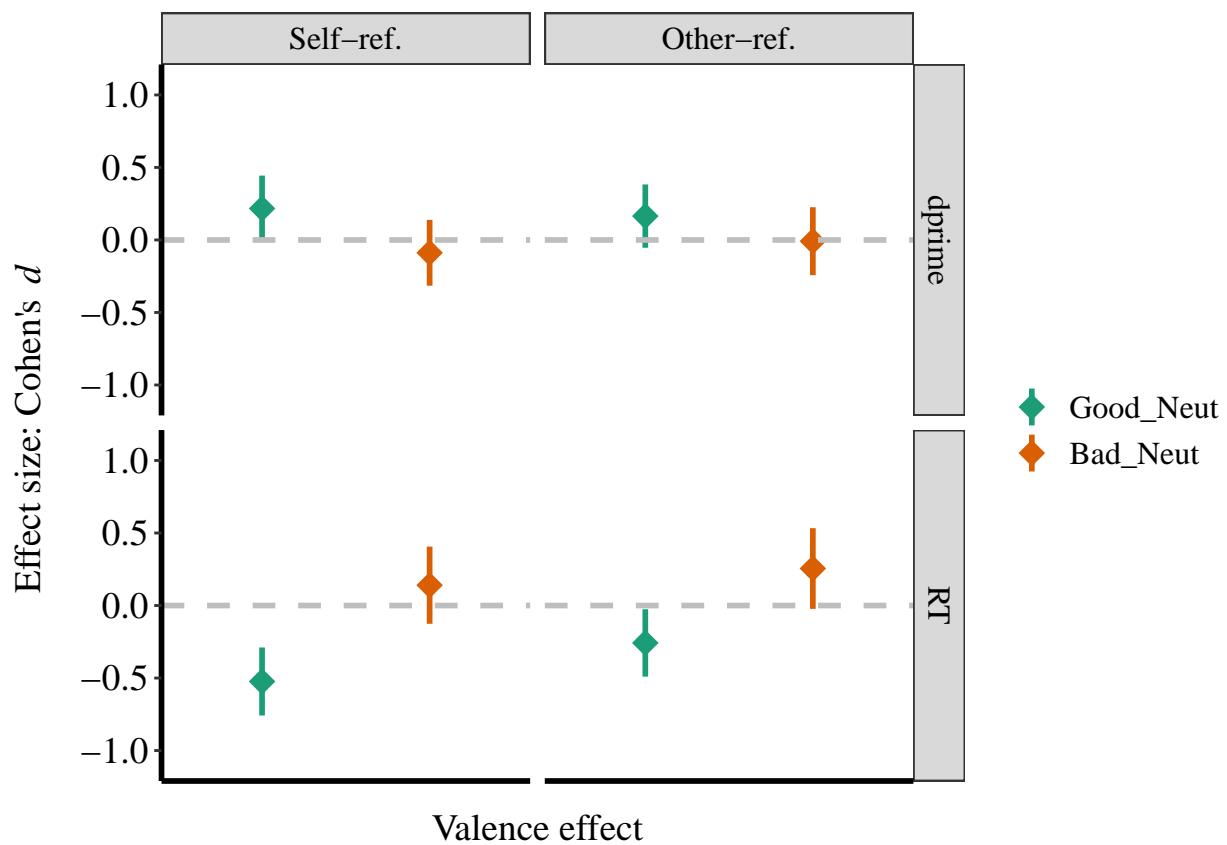


Figure 33. Effect size (Cohen's  $d$ ) of Valence in Exp4b.

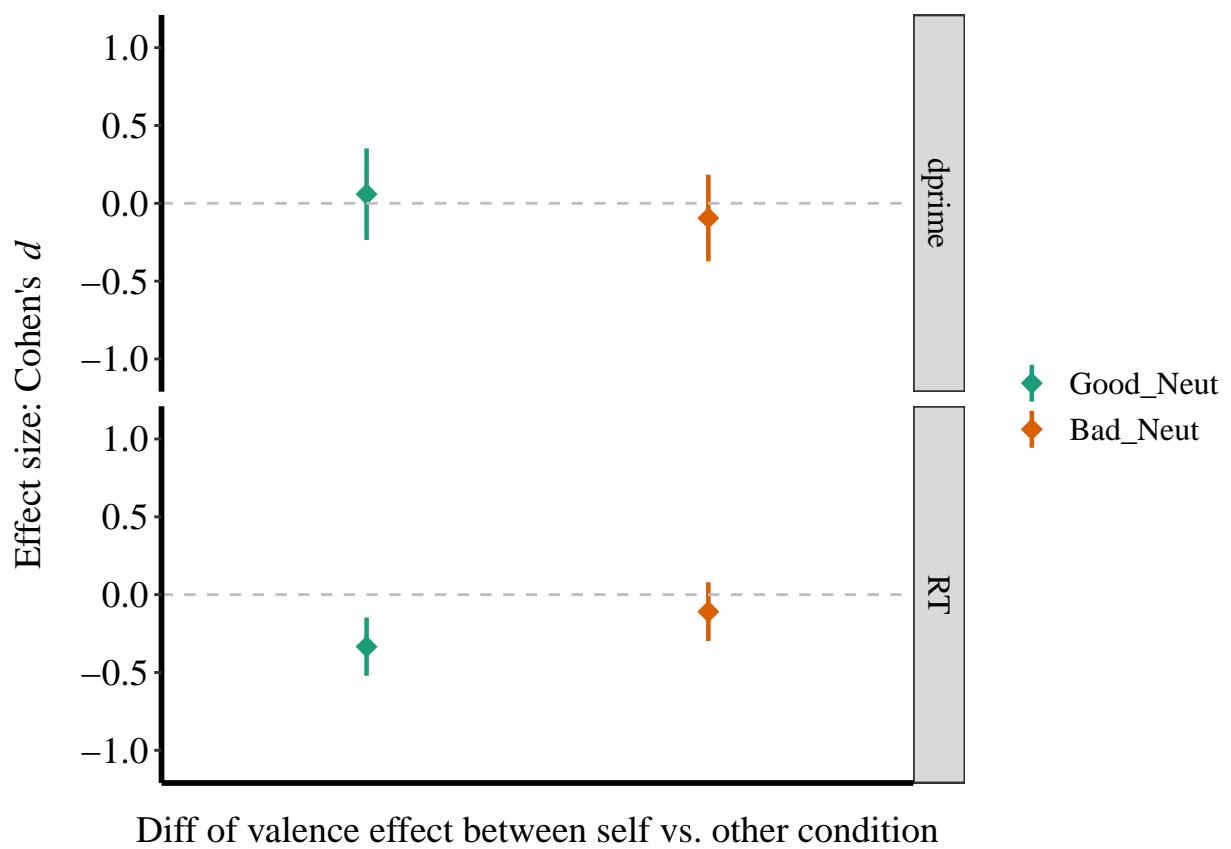


Figure 34. Effect size (Cohen's  $d$ ) of Valence in Exp4b.

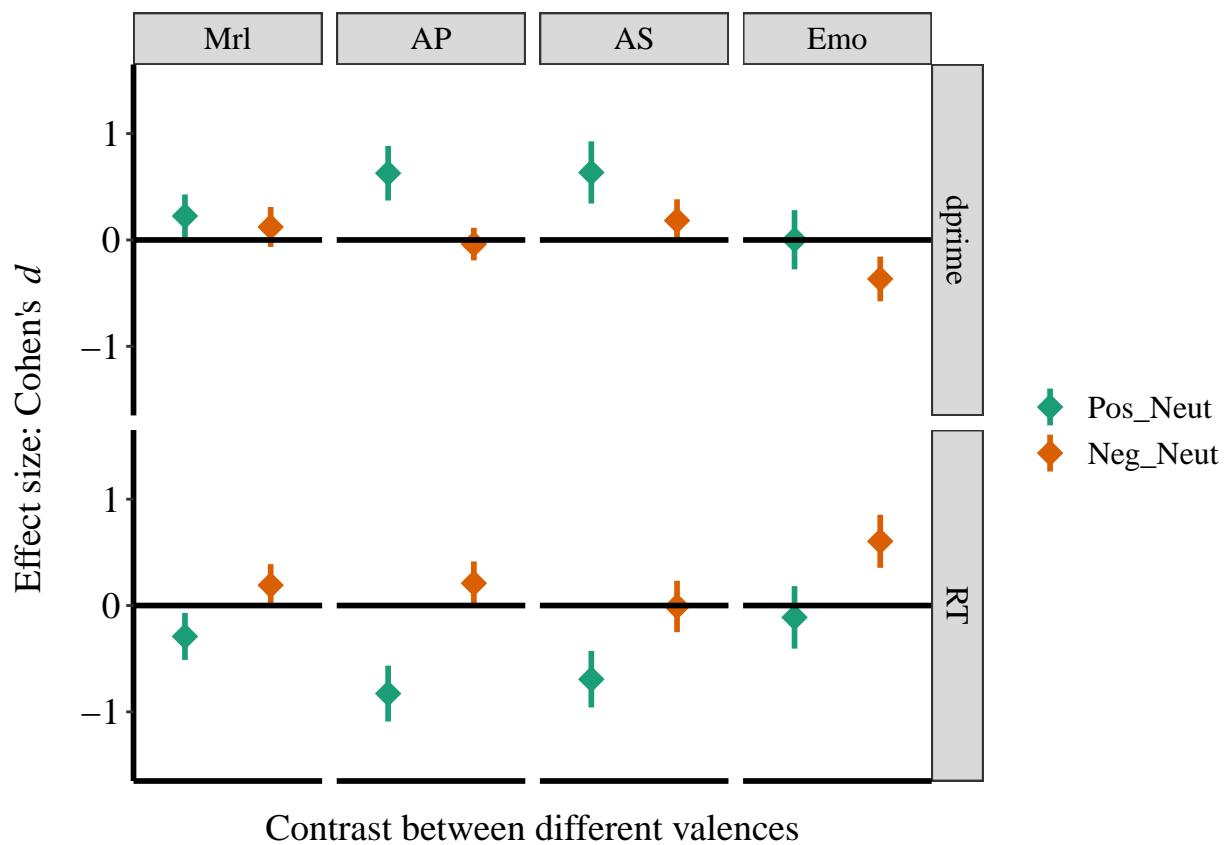


Figure 35. Effect size (Cohen's  $d$ ) of Valence in Exp5.

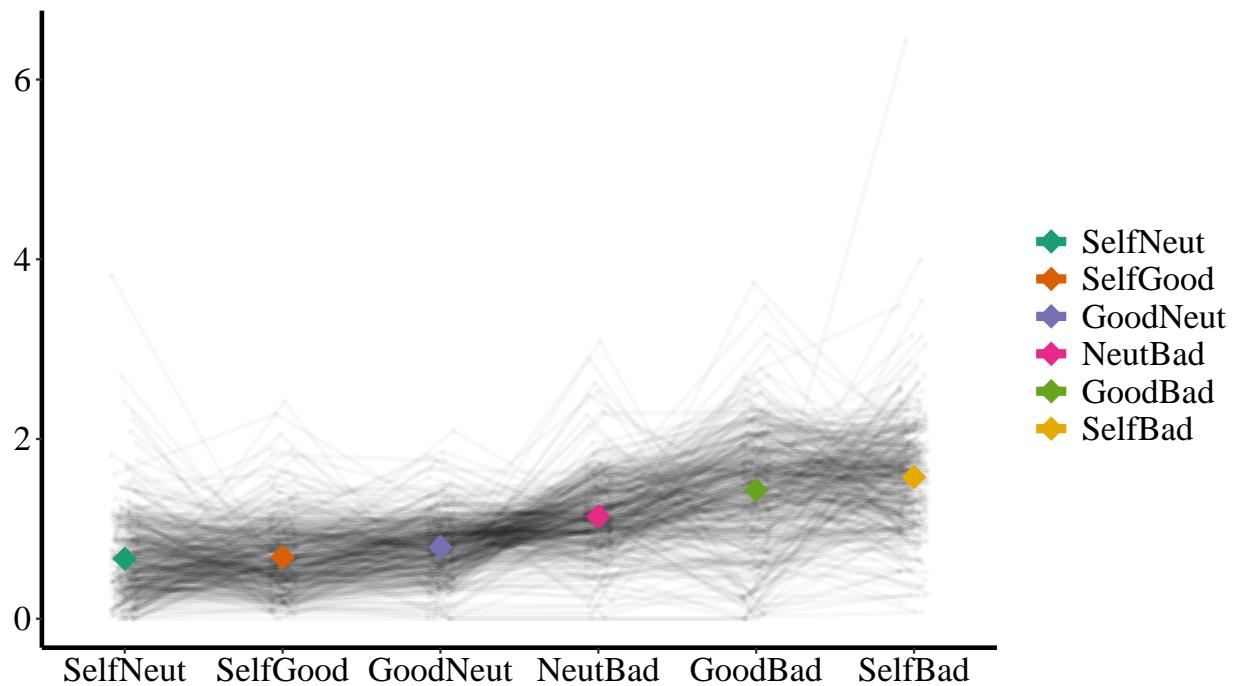


Figure 36. Self-rated personal distance

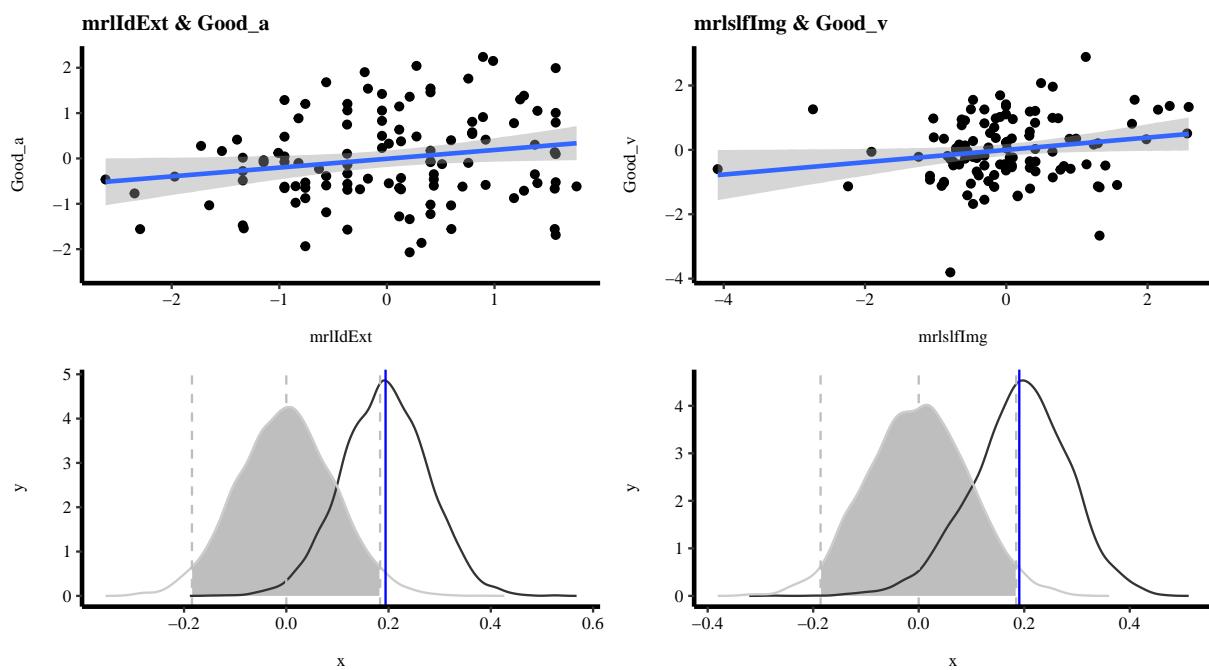
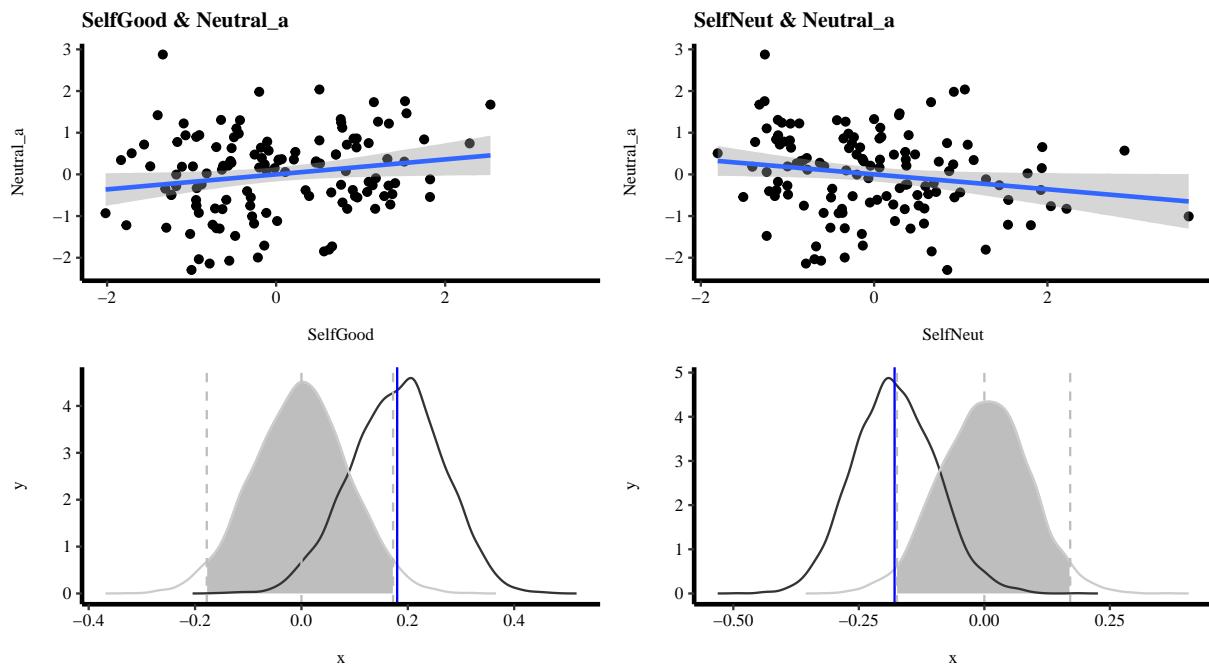
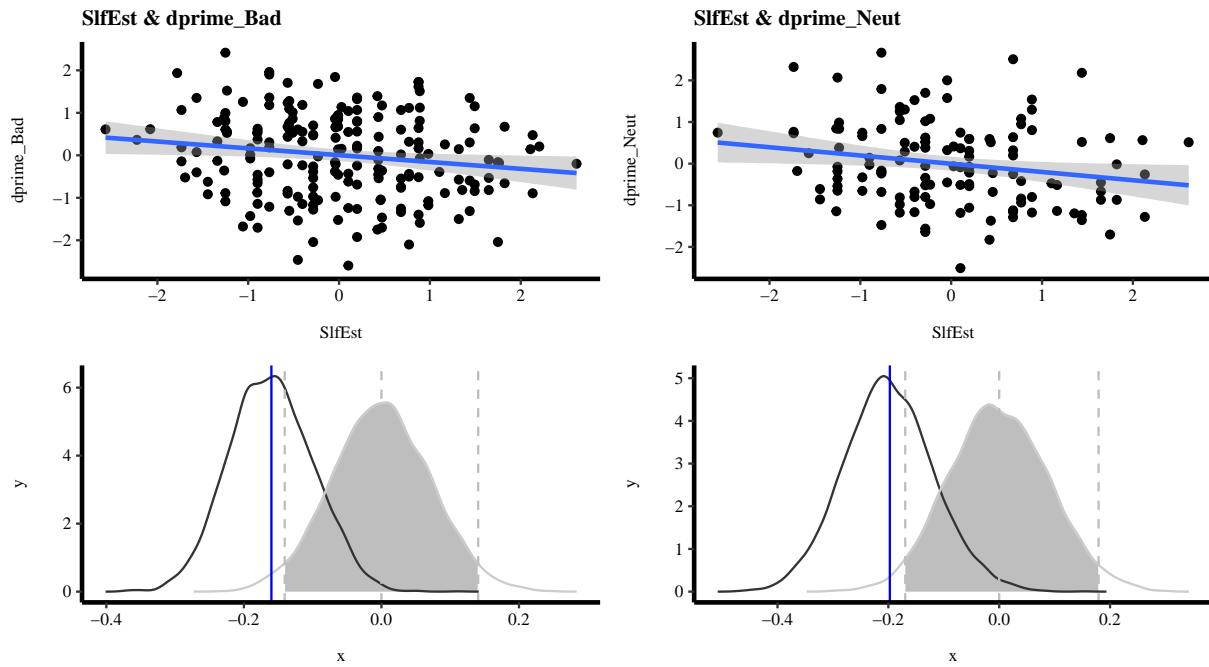


Figure 37. Correlation between moral identity and boundary separation of good condition; moral self-image and drift rate of good condition



*Figure 38.* Correlation between personal distance and boundary separation of neutral condition



*Figure 39.* Correlation between self esteem and d prime of bad and neutral conditions