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

Hu Chuan-Peng[1,2], Kaiping Peng[3], & Jie Sui[3,4]

[1] TBA

[2] Leibniz Institute for Resilience Research, 55131 Mainz, Germany

[3] Tsinghua University, 100084 Beijing, China

[4] University of Aberdeen, Aberdeen, Scotland

Author Note

Hu Chuan-Peng, Leibniz Institute for Resilience Research (LIR). Kaiping Peng, Department of Psychology, Tsinghua University, 100084 Beijing, China. Jie Sui, School of Psychology, University of Aberdeen, Aberdeen, Scotland.

Authors contriubtion: HCP, JS, & KP design the study, HCP collected the data, HCP analyzed the data and drafted the manuscript. KP & JS supported this project.

Correspondence concerning this article should be addressed to Hu Chuan-Peng, Langenbeckstr. 1, Neuroimaging Center, University Medical Center Mainz, 55131 Mainz, Germany. E-mail: hcp4715@gmail.com

Abstract

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

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

Word count: X

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

Introduction

XXXX In perceptual matching, same is faster than different (Farell, 1985; Krueger, 1978). Automatic processing (Spruyt & Houwer, 2017)

Van Zandt, Colonius, and Proctor (2000): A comparison of two response time models applied to perceptual matching

Yakushijin, ReikoJacobs, Robert A (2020), Are People Successful at Learning Sequential Decisions on a Perceptual Matching Task?

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. Psychological Review, 108(1), 257–272. https://doi.org/10.1037/0033-295X.108.1.257

We reported results from eleven experiments. In first set of experiments, we found that shapes associated with morally positive person label were responded faster and more accurately. In the second set of experiment, we explore the potential role of good self in perceptual matching task and added one more independent variable, we found that the effect was mainly on good self. In the third part we tested whether the morality will automatically binds with person-relevance. Finally, we explore the correlation between behavioral task and questionnaire scores.

Disclosures

We reported all the measurements, analyses, and results in all the experiments in the current study. Participants whose overall accuracy lower than 60% were excluded from analysis. Also, the accurate responses with less than 200ms reaction times were excluded from the analysis. To have a better overview of the effect reported in this series

experiment, we reported the synthesized results in the main text and individual experiment in supplementary materials.

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

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

Part 1: Moral valence effect

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

Experiment 1a

Methods.

Participants.

57 college students (38 female, age = 20.75 $\pm$ 2.54 years) participated. 39 of them were recruited from Tsinghua University community in 2014; 18 were recruited from Wenzhou University in 2017. All participants were right-handed except one, and all had normal or corrected-to-normal vision. Informed consent was obtained from all participants

prior to the experiment according to procedures approved by the local ethics committees. 6 participant's data were excluded from analysis because nearly random level of accuracy, leaving 51 participants (34 female, age = 20.72 ± 2.44 years).

Stimuli and Tasks.

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

Procedure.

This experiment had two phases. First, there was a brief learning stage. Participants were asked to learn the relationship between geometric shapes (triangle, square, and circle) and different person (bad person, a good person, or a neutral person). For example, a participant was told, "bad person is a circle; good person is a triangle; and a neutral person is represented by a square." After participant remember the associations (usually in a few minutes), participants started a practicing phase of matching task which has the exact task as in the experimental task. In the experimental task, participants judged whether shape–label pairs, which were subsequently presented, were correct. Each trial started with the presentation of a central fixation cross for 500 ms. Subsequently, a pairing of a shape and label (good person, bad person, and neutral person) was presented for 100 ms. The pair presented could confirm to the verbal instruction for each pairing given in the training stage, or it could be a recombination of a shape with a different label, with the shape–label pairings being generated at random. The next frame showed a blank for 1100ms. Participants were expected to judge whether the shape was correctly assigned to the person by pressing one of the two response buttons as quickly and accurately as possible within this timeframe (to encourage immediate responding). Feedback (correct or incorrect) was given on the screen for 500 ms at the end of each trial, if no response detected, "too slow" was presented to remind participants to accelerate. Participants were informed of their

overall accuracy at the end of each block. The practice phase finished and the experimental task began after the overall performance of accuracy during practice phase achieved 60%. For participants from the Tsinghua community, they completed 6 experimental blocks of 60 trials. Thus, there were 60 trials in each condition (bad-person match, bad-person nonmatch, good-person match, good-person nonmatch, neutral-person match, and neutral-person nonmatch). For the participants from Wenzhou University, they finished 6 blocks of 120 trials, therefore, 120 trials for each condition.

Data analysis.

We analyzed accuracy performance using a signal detection theory approach. The performance in each match condition was combined with that in the nonmatch condition with the same shape to form a measure of d', the match trials were regarded as signal while the nonmatch trials were regarded as noise (Sui, He, & Humphreys, 2012). Given that the match and nonmatch trials are presented in the same way and had same number of trials across all studies, we assume that participants' inner distribution of these two types of trials had equal variance but may had different means. That is, we used the equal variance Gaussian SDT model (EVSDT) here (Rouder & Lu, 2005). **Trials without response were excluded from the analysis.**

We analyzed the d prime and reaction times using the generalized linear model (GLM). The GLM approach didn't assume a particular generative model of the data but using linear model to fit the data. We applied the traditional repeated measures ANOVA, which is a special case of GLM. We also reported results from Bayesian generalized linear model approach.

We also analyzed the accuracy and reaction times data using the drift-diffusion model (DDM), which combine the accuracy and reaction times together.

Classic NHST.

In the classic NHST approach, we've used the maximum likelihood estimate of the

EVSDT parameters (mainly, the sensitivity d') separately for each participant in the each experiment condition (Hu et al., 2020; Sui et al., 2012), and then used the parameter of each condition from each participant as dependent variable for repeated measures ANOVA. The d' is calculated as the difference of the standardized hit and false alarm rats (Stanislaw & Todorov, 1999):

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

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

For the RTs, we used mean RTs of the accurate trials of each condition and subject for repeated measures ANOVA.

The repeated measure ANOVA was done by `afex` package.

Bayesian hierarchical generalized linear model (GLM).

The maximum likelihood estimate of parameters of EVSDT may ignored the uncertainty in estimates of the parameters (Rouder & Lu, 2005). Bayesian generalized linear model (GLM) can help to estimate the uncertainty of parameters. We used BRMs (Bürkner, 2017; Carpenter et al., 2017) to model the data.

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

$$y_i \sim Bernoulli(p_i)$$

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

$$\Phi(p_i) = \beta_0 + \beta_1 IsMatch_i$$

The outcomes y_i are 0 if the participant responded "nonmatch" on trial i, 1 if they responded "match". The probability of the "match" response for trial i for a participant is p_i. We then write the generalized linear model on the probits (z-scores; Φ, "Phi") of ps. Φ is the cumulative normal density function and maps z scores to probabilities. Given this parameterization, the intercept of the model (β_0) is the standardized false alarm rate (probability of saying 1 when predictor is 0), which we take as our criterion c. The slope of the model (β_1) is the increase of saying 1 when predictor is 1, in z-scores, which is another expression of d'. Therefore, c = -zHR = $-\beta_0$, and $d' = \beta_1$.

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

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

Similarly, the generalized linear model was extended to two levels:

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

The outcomes y_{ij} are 0 if participant j responded "nonmatch" on trial i, 1 if they responded "match". The probability of the "match" response for trial i for subject j is p_{ij}. We again can write the generalized linear model on the probits (z-scores; Φ, "Phi") of ps.

The subjective-specific intercepts ($\beta_0 = -zFAR$) and slopes ($\beta_1 = d'$) are describe

by multivariate normal with means and a covariance matrix for the parameters.

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

In the same vein, when trying to estimate the parameter across different experiments, we can further consider the participant is nested in different experiments, and each experiment, because of its variation in stimuli or stimuli presentation, may have different intercepts and slopes. In this case, we can use a nested hierarchical model to model all the experiment with similar design:

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

where

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

The outcomes y_{ijk} are 0 if participant j in experiment k responded "nonmatch" on trial i, 1 if they responded "match".

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

and the experiment level parameter mu_{0k} and mu_{1k} is from a higher order distribution:

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

in which μ_0 and μ_1 means the population level parameter.

Using the Bayesian hierarchical model, we can directly estimate the over-all effect of valence on d' across all experiments with similar experimental design, instead of using a two-step approach where we first estimate the d' for each participant and then use a random effect model meta-analysis (Goh, Hall, & Rosenthal, 2016).

For the reaction time, there are many criticism about using the mean RTs as the representation of each participant (Rousselet & Wilcox, 2019), to better capture a representative parameter for RTs, we used the log normal distribution (https://lindeloev.github.io/shiny-rt/#34_(shifted)_log-normal). This distribution has two parameters: μ, σ. μ is the mean of the logNormal distribution, and σ is the disperse of the distribution. The log normal distribution can be extended to shifted log normal distribution, with one more parameter: shift, which is the earliest possible response.

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

Shifted log-normal distribution:

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

y_{ij} is the RT of the ith trial of the jth participants.

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

$$\sigma_j \sim Cauchy()$$

This model can be easily expand to three-level model in which participants and experiments are two group level variable and participants were nested in the experiments.

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

y_{ijk} is the RT of the ith trial of the jth participants in the kth experiment.

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

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

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

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

Hierarchical drift diffusion model (HDDM).

To further explore the psychological mechanism under perceptual decision-making, we used HDDM (Wiecki, Sofer, & Frank, 2013) to model our RTs and accuracy data. We used the prior implemented in HDDM, that is, informative priors that constrains parameter estimates to be in the range of plausible values based on past literature (Matzke & Wagenmakers, 2009)

Results.

Classic analytical approach.

d *prime.*

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

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

Reaction times.

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

We then carried out two separate ANOVA for Match and Mismatched trials. For matched trials, we found the effect of valence . We further examined the effect of valence for both self and other for matched trials. We found that shapes associated with Good Person (684 ms, SE = 11.5) responded faster than Neutral (709 ms, SE = 11.5), $t(50)$ = -2.265, $p = 0.0702$) and Bad Person (728 ms, SE = 11.7), $t(50)$ = -4.41, $p = 0.0002$), and the Neutral condition was faster than the Bad condition, $t(50)$ = -2.495, $p = 0.0415$). For non-matched trials, there was no significant effect of Valence ().

Bayesian hierarchical GLM.

d prime.

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

Interesting, we also found the criteria for three conditions also differ, the shapes tagged with good person has the highest criteria (-1.01, [-1.14 -0.88]), followed by shapes tagged with neutral person(-1.06, [-1.21 -0.92]), and then the shapes tagged with bad

person(-1.11, [-1.25 -0.97]). However, pair-wise comparison showed that only showed strong evidence for the difference between good and bad conditions.

Reaction times.

We fitted a Bayesian hierarchical GLM for RTs, with a log-normal distribution as the link function. We used the posterior distribution of the regression coefficient to make statistical inferences. As in previous studies, the matched conditions are much faster than the mismatched trials ($P_{PosteriorComparison} = 1$). We focused on matched trials only, and compared different conditions: Good is faster than the neutral, $P_{PosteriorComparison} = .99$, it was also faster than the Bad condition, $P_{PosteriorComparison} = 1$. And the neutral condition is faster than the bad condition, $P_{PosteriorComparison} = .99$. However, the mismatched trials are largely overlapped. See Figure 2.

HDDM.

We fitted our data with HDDM, using the response-coding (See also, 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 3).

Design and Procedure

This series of experiments started to test the effect of instantly acquired moral valence on perceptual decision-making. For this purpose, we used the social associative learning paradigm (or tagging paradigm)(Sui et al., 2012), in which participants first learned the associations between geometric shapes and labels of person with different moral valence (e.g., in first three studies, the triangle, square, and circle and good person, neutral

person, and bad person, respectively). The associations of the shapes and label were counterbalanced across participants. After remembered the associations, participants finished a practice phase to familiar with the task, in which they viewed one of the shapes upon the fixation while one of the labels below the fixation and judged whether the shape and the label matched the association they learned. When participants reached 60% or higher accuracy at the end of the practicing session, they started the experimental task which was the same as in the practice phase.

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

For experiment 3a, 3b, 4a, 4b, 6b, 7a, and 7b, we included self-reference as another within-subject variable in the experimental design. For example, the experiment 3a directly extend the design of experiment 1a into a 2 (matchness: match vs. nonmatch) by 2 (reference: self vs. other) by 3 (moral valence: good vs. neutral vs. bad) within-subject design. Thus in experiment 3a, there were six conditions (good-self, neutral-self, bad-self, good-other, neutral-other, and bad-other) and six shapes (triangle, square, circle, diamond,

pentagon, and trapezoids). The experiment 6b was an EEG experiment extended from experiment 3a but presented the label and shape sequentially. Because of the relatively high working memory load (six label-shape pairs), experiment 6b were conducted in two days: the first day participants finished perceptual matching task as a practice, and the second day, they finished the task again while the EEG signals were recorded. Experiment 3b was designed to separate the self-referential trials and other-referential trials. That is, participants finished two different blocks: in the self-referential blocks, they only responded to good-self, neutral-self, and bad-self, with half match trials and half non-match trials; for the other-reference blocks, they only responded to good-other, neutral-other, and bad-other. Experiment 7a and 7b were designed to test the cross task robustness of the effect we observed in the aforementioned experiments (see, Hu et al., 2020). The matching task in these two experiments shared the same design with experiment 3a, but only with two moral valence, i.e., good vs. bad. We didn't include the neutral condition in experiment 7a and 7b because we found that the neutral and bad conditions constantly showed non-significant results in experiment $1 \sim 6$.

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

Finally, experiment 5 was design to test the specificity of the moral valence. We extended experiment 1a with an additional independent variable: domains of the valence words. More specifically, besides the moral valence, we also added valence from other domains: appearance of person (beautiful, neutral, ugly), appearance of a scene (beautiful,

neutral, ugly), and emotion (happy, neutral, and sad). Label-shape pairs from different domains were separated into different blocks.

E-prime 2.0 was used for presenting stimuli and collecting behavioral responses, except that experiment 7a and 7b used Matlab Psychtoolbox (Brainard, 1997; Pelli, 1997). For participants recruited in Tsinghua University, they finished the experiment individually 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° of visual angle. For participants recruited in Wenzhou University, they finished the experiment in a group consisted of $3 \sim 12$ 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.

Data analysis

Analysis of individual study. The individual experiment's results were reported in supplementary materials. We used the `tidyverse` of r (see script `Load_save_data.r`) to exclude the practicing trials, invalid trials of each participants, and invalid participants, if

there were any, in the raw data.

Results of each experiment were analyzed as in Sui et al. (2012). That is, the accuracy performance using a signal detection approach, in which the performance in each match condition was combined with that in the nonmatch condition with the same shape to form a measure of d' . Trials without response were coded either as "miss" (match trials) or "false alarm" (nonmatch trials). For the reaction times (RTs), only RTs of accurate trials were analyzed.

Both signal detection theory analysis of accuracy and RTs were analyzed in Frequentists' approach and Bayesian approach. In the Frequentists' approach, we calculated the d' using Maximum Likelihood approach and then subjected the d' estimated from each participant to repeated measures analyses of variance (repeated measures ANOVA), via `afex` []; we used the mean RTs of each participant in each condition and subject these mean value to repeated measures ANOVA too. We reported the results from significance test and effect sizes (including 95% confidence intervals). To control the false positive rate when conducting the post-hoc comparisons, we used Bonferroni correction.

In the Bayesian approach, we used `brms` (Bürkner, 2017; Carpenter et al., 2017) to implement the Bayesian hierarchical generalized linear model to estimate the effect of valence and self-referential. See supplementary materials for the results of each experiment's method and results.

Finally, we also explored the psychological processes during the perceptual decision-making using the drift-diffusion model (DDM). We used HDDM (Wiecki et al., 2013) for this purpose.

Synthesized results. We reported the synthesized results from the experiments, because many of them shared the similar experimental design. We reported the results in five parts: valence effect, explicit interaction between valence and self-relevance, implicit interaction between valence and self-relevance, specificity of valence effect, and

behavior-questionnaire correlation.

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

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

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

M_1 is the mean of the first condition, sd_1 is the standard deviation of the first condition, while M_2 is the mean of the second condition, sd_2 is the standard deviation of the second condition. r is the correlation coefficient between data from first and second condition. n is the number of data point (in our case the number of participants included in our research).

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

Valence effect. We synthesized effect size of d' and RT from experiment 1a, 1b, 1c, 2, 5 and 6a for the valence effect. We reported the synthesized the effect across all

experiments that tested the valence effect, using the mini meta-analysis approach (Goh et al., 2016).

Explicit interaction between Valence and self-relevance. The results from experiment 3a, 3b, 6b, 7a, and 7b. These experiments explicitly included both moral valence and self-reference.

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

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

Behavior-Questionnaire correlation. Finally, we explored correlation between results from behavioral results and self-reported measures.

For the questionnaire part, we are most interested in the self-rated distance between different person and self-evaluation related questionnaires: self-esteem, moral-self identity,

and moral self-image. Other questionnaires (e.g., personality) were not planned to correlated with behavioral data were not included. Note that all data were reported in (Liu et al., 2020).

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

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

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

The DDM analyses were finished by HDDM, as reported in Hu et al. (2020). That is, we used the response code approach, match response were coded as 1 and nonmatch responses were coded as 0. To fully explore all parameters, we allow all four parameters of DDM free to vary. We then extracted the estimation of all the four parameters for each participants for the correlation analyses. However, because the starting point is only related to response (match vs. non-match) but not the valence of the stimuli, we didn't included it in correlation analysis.

We used Pearson correlation to quantify the correlation. For those correlation that is significant ($p < 0.05$), we further tested the robustness of the correlation using bootstrap by `BootES` package (Kirby & Gerlanc, 2013). To avoid false positive, we further determined the threshold for signifcant by permutation. More specifically, for each pairs that initially with $p < .05$, we randomly shuffle the participants data of each score and calculated the

correlation between the shuffled vectors. After repeating this procedure for 5000 times, we choose arrange these 5000 correlation coefficients and use the 95% percentile number as our threshold.

Results

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 4 left panel.

Interaction between valence and self-reference

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

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

For the self-referential condition, we found the same pattern as in the first part of results. That is we found significant differences between positive and neutral as well as positive and negative, but not neutral and negative. The effect size of RT between positive and negative is Cohen's d = -0.89 ± 0.12, 95% CI [-1.11 -0.66]; on *d'* was Cohen's d = 0.61 ± 0.09, 95% CI [0.44 0.78]. The effect was also observed between positive and neutral condition, RT: Cohen's d = -0.76 ± 0.13, 95% CI [-1.01 -0.50]; *d'*: Cohen's d = 0.69 ± 0.14, 95% CI [0.42 0.96]. And the difference between neutral and bad conditions are not significant, RT: Cohen's d = 0.03 ± 0.13, 95% CI [-0.22 0.29]; *d'*: Cohen's d = 0.08 ± 0.08, 95% CI [-0.07 0.24]. See Figure 4 the middle panel.

For the other-referential condition, we found that only the difference between positive and negative on RT was significant, all the other conditions were not. The effect size of RT between positive and negative is Cohen's d = -0.28 ± 0.05, 95% CI [-0.38 -0.17]; on *d'* was Cohen's d = -0.02 ± 0.08, 95% CI [-0.17 0.13]. The effect was not observed between positive and neutral condition, RT: Cohen's d = -0.12 ± 0.10, 95% CI [-0.31 0.06]; *d'*: Cohen's d = 0.01 ± 0.08, 95% CI [-0.16 0.17]. And the difference between neutral and bad conditions are not significant, RT: Cohen's d = 0.14 ± 0.09, 95% CI [-0.03 0.31]; *d'*: Cohen's d = 0.05 ± 0.07, 95% CI [-0.08 0.18]. See Figure 4 right panel.

Generalizibility of the valence effect

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

For experiment 4a, when self-reference was the target and moral valence was task-irrelevant, we found that only under the implicit self-referential condition, i.e., when the moral words were presented as task irrelevant stimuli, there was the main effect of valence and interaction between valence and reference for both d prime and RT (See supplementary results for the detailed statistics). For d prime, we found good-self

condition (2.55 ± 0.86) had higher d prime than bad-self condition (2.38 ± 0.80); good self condition was also higher than neutral self (2.45 ± 0.78) but there was not statistically significant, while the neutral-self condition was higher than bad self condition and not significant neither. For reaction times, good-self condition (654.26 ± 67.09) were faster relative to bad-self condition (665.64 ± 64.59), and over neutral-self condition (664.26 ± 64.71). The difference between neutral-self and bad-self conditions were not significant. However, for the other-referential condition, there was no significant differences between different valence conditions. See Figure 5.

For experiment 4b, when valence was the target and the identity was task-irrelevant, we found a strong valence effect (see supplementary results and Figure 6, Figure 7).

In this experiment, the advantage of good-self condition can only be disentangled by comparing the self-referential and other-referential conditions. Therefore, we calculated the differences between the valence effect under self-referential and other referential conditions and used the weighted variance as the variance of this differences. We found this modulation effect on RT. The valence effect of RT was stronger in self-referential than other-referential for the Good vs. Neutral condition (-0.33 ± 0.01), and to a less extent the Good vs. Bad condition (-0.17 ± 0.01). While the size of the other effect's CI included zero, suggestion those effects didn't differ from zero. See Figure 8.

Specificity of valence effect

In this part, we analyzed the results from experiment 5, which included positive, neutral, and negative valence from four different domains: morality, emotion, aesthetics of human, and aesthetics of scene. We found interaction between valence and domain for both d prime and RT (match trials). A common pattern appeared in all four domains: each domain showed a binary results instead of gradient on both d prime and RT. For morality, aesthetics of human, and aesthetics of scene, the positive conditions had advantages over

both neutral and negative conditions (greater d prime and faster RT), and neutral and negative conditions didn't differ from each other. But for the emotional stimuli, it was the positive and neutral had advantage over negative conditions, while positive and neutral conditions were not significantly different. See supplementary materials for detailed statistics. Also note that the effect size in moral domain is smaller than the aesthetic domains (beauty of people and beauty of scene). See Figure 9.

Self-reported personal distance

See Figure 10.

Correlation analyses

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

We focused on the task-questionnaire correlation, the results revealed that the score from three questionnaire are related to behavioral responses data. First, the external moral identity is positively correlated with boundary separation of moral good condition, $r = 0.194$, 95% CI [0.023 0.350]); the moral self image is positively correlated with the drift rate ($r = 0.191$, 95% CI [-0.016 0.354]) of the morally good condition. See Figure 11.

Second, we found the personal distance between self and good is positively correlated with the boundary separation of neutral condition and the self-neutral distance is negatively correlated with the boundary separation of neutral condition. See figure 12

Third, we found the self esteem score was negative correlated with the d' of bad conditions ($r = -0.16$, 95% CI [-0.277 -0.038]) and the neutral conditions ($r = -.197$, 95% CI [-0.348 -0.026]). See Figure 13.

We also explored the correlation between behavioral data and questionnaire scores separately for experiments with and without self-referential. For experiments without

self-referential (Valence effect), we found the personal distance between Good-person and self is positively correlated with boundary separation of good conditions, r = 0.292, 95% [0.071 0.485]. also personal distance between the bad and neutral person is positively correlated with non-responding time of bad and neutral conditions, r = 0.249, 0.233, respectively.

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

Discussion

References

Anderson, E., Siegel, E. H., Bliss-Moreau, E., & Barrett, L. F. (2011). The visual impact of gossip. *Science*, *332*(6036), 1446–1448. https://doi.org/10.1126/science.1201574

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. Journal Article.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software; Vol 1, Issue 1 (2017).* Journal Article. Retrieved from https://www.jstatsoft.org/v080/i01%0Ahttp://dx.doi.org/10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). Journal Article. https://doi.org/10.18637/jss.v076.i01

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). Book, New York: Sage.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186–205. Journal Article. https://doi.org/10.1037/1082-989X.3.2.186

Farell, B. (1985). "Same"–"different" judgments: A review of current controversies in perceptual comparisons. *Psychological Bulletin*, *98*(3), 419–456. Journal Article. https://doi.org/10.1037/0033-2909.98.3.419

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*(1), 22–29. https://doi.org/10.1016/j.cognition.2014.02.007

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*(10), 535–549. Journal Article. https://doi.org/10.1111/spc3.12267

Hu, C.-P., Lan, Y., Macrae, C. N., & Sui, J. (2020). Good me bad me: Does valence influence self-prioritization during perceptual decision-making? *Collabra: Psychology*, *6*(1), 20. Journal Article. https://doi.org/10.1525/collabra.301

Kirby, K. N., & Gerlanc, D. (2013). BootES: An r package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, *45*(4), 905–927. https://doi.org/10.3758/s13428-013-0330-5

Krueger, L. E. (1978). A theory of perceptual matching. *Psychological Review*, *85*(4), 278–304. Journal Article. https://doi.org/10.1037/0033-295X.85.4.278

Liu, Q., Wang, F., Yan, W., Peng, K., Sui, J., & Hu, C.-P. (2020). Questionnaire data from the revision of a chinese version of free will and determinism plus scale. *Journal of Open Psychology Data*, *8*(1), 1. Journal Article. https://doi.org/10.5334/jopd.49/

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*(5), 798–817. https://doi.org/10.3758/PBR.16.5.798

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. Journal Article.

Rouder, J. N., & Lu, J. (2005). An introduction to bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604. Journal Article. https://doi.org/10.3758/bf03196750

Rousselet, G. A., & Wilcox, R. R. (2019). Reaction times and other skewed distributions: Problems with the mean and the median. *Meta-Psychology*. preprint. https://doi.org/10.1101/383935

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after p-hacking. Conference Proceedings. https://doi.org/10.2139/ssrn.2205186

Spruyt, A., & Houwer, J. D. (2017). On the automaticity of relational stimulus processing: The (extrinsic) relational simon task. *PLoS One*, *12*(10), e0186606. Journal Article. https://doi.org/10.1371/journal.pone.0186606

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. Journal Article. https://doi.org/10.3758/BF03207704

Sui, J., He, X., & Humphreys, G. W. (2012). Perceptual effects of social salience: Evidence from self-prioritization effects on perceptual matching. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(5), 1105–1117. Journal Article. https://doi.org/10.1037/a0029792

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*(2), 208–256. https://doi.org/10.3758/BF03212980

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in Neuroinformatics*, *7*.

https://doi.org/10.3389/fninf.2013.00014

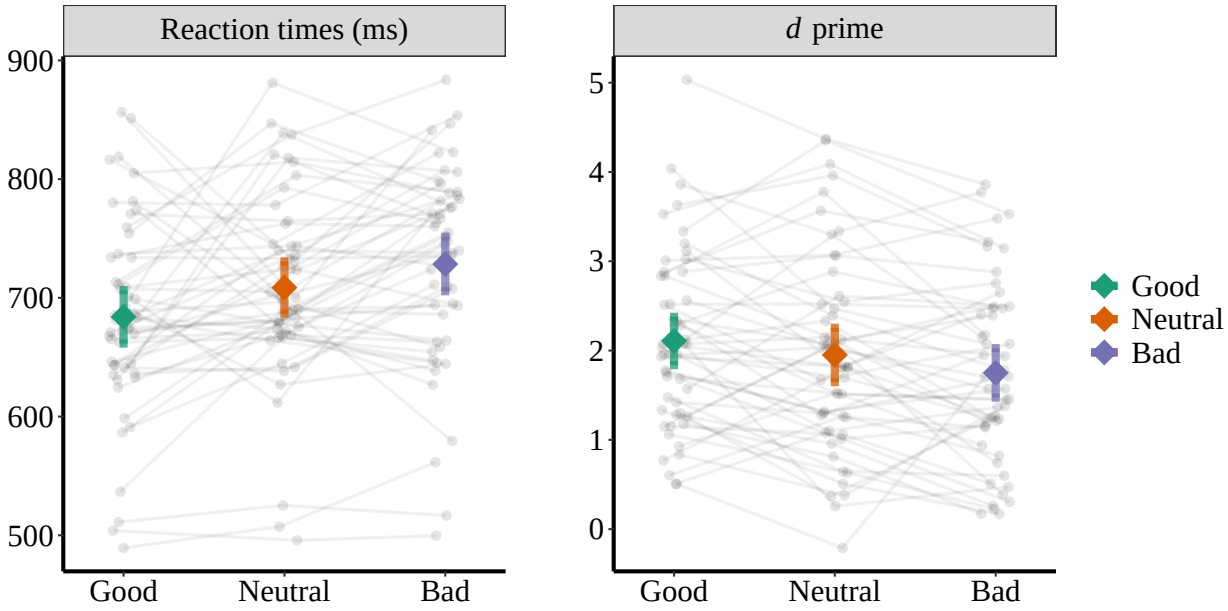

Figure 1. RT and *d* prime of Experiment 1a.

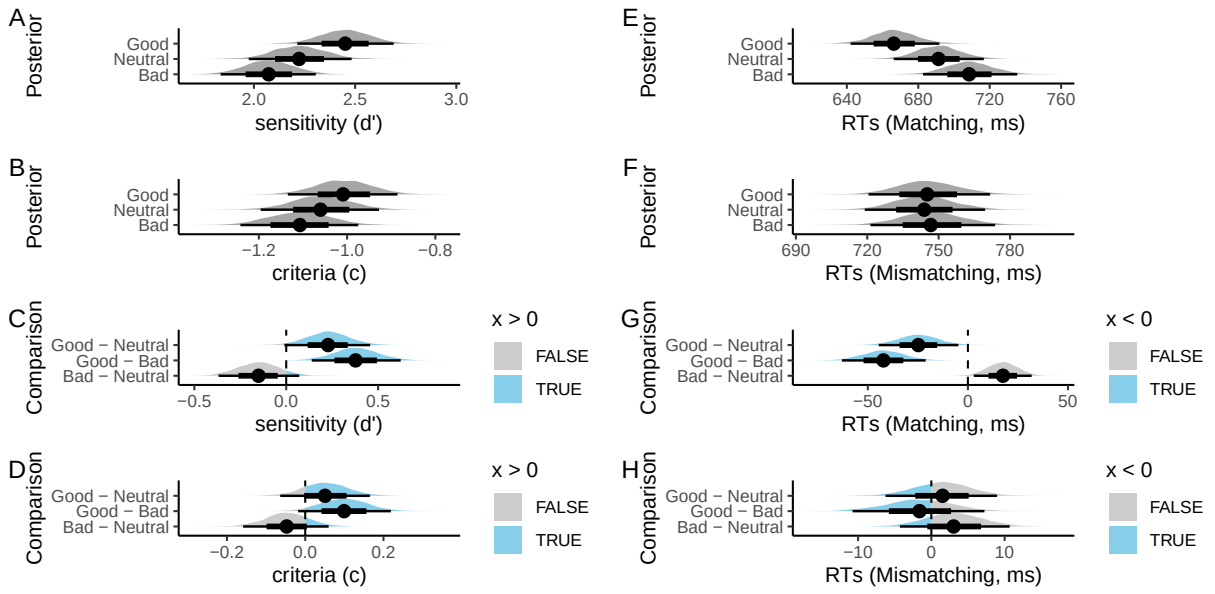

Figure 2. Exp1a: Results of Bayesian GLM analysis.

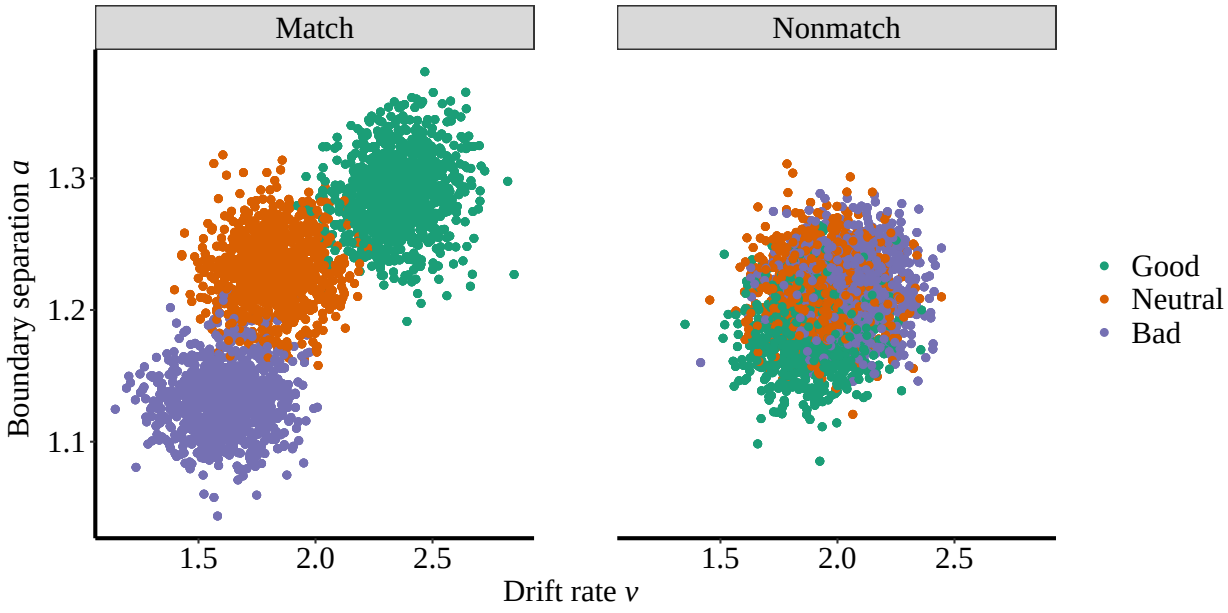

Figure 3. Exp1a: Results of HDDM.

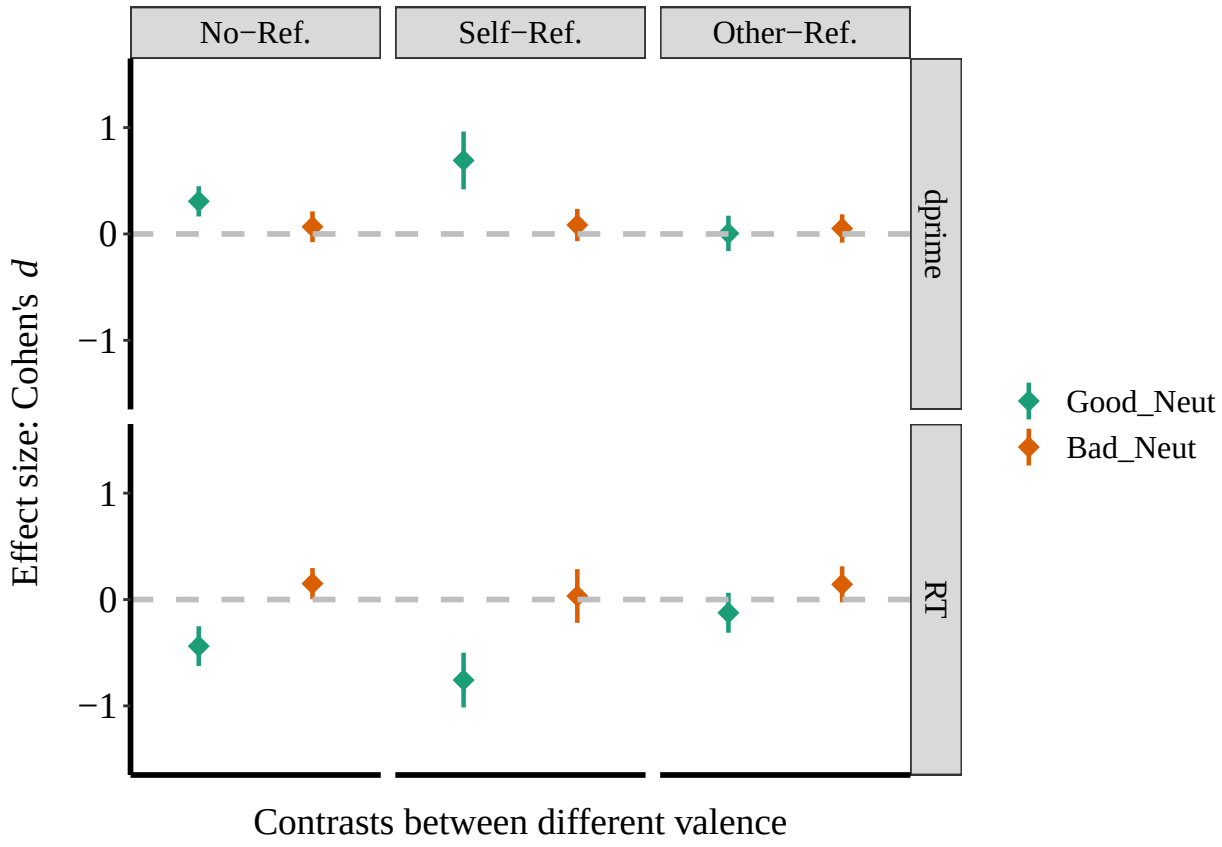

Figure 4. Effect size (Cohen's *d*) of Valence.

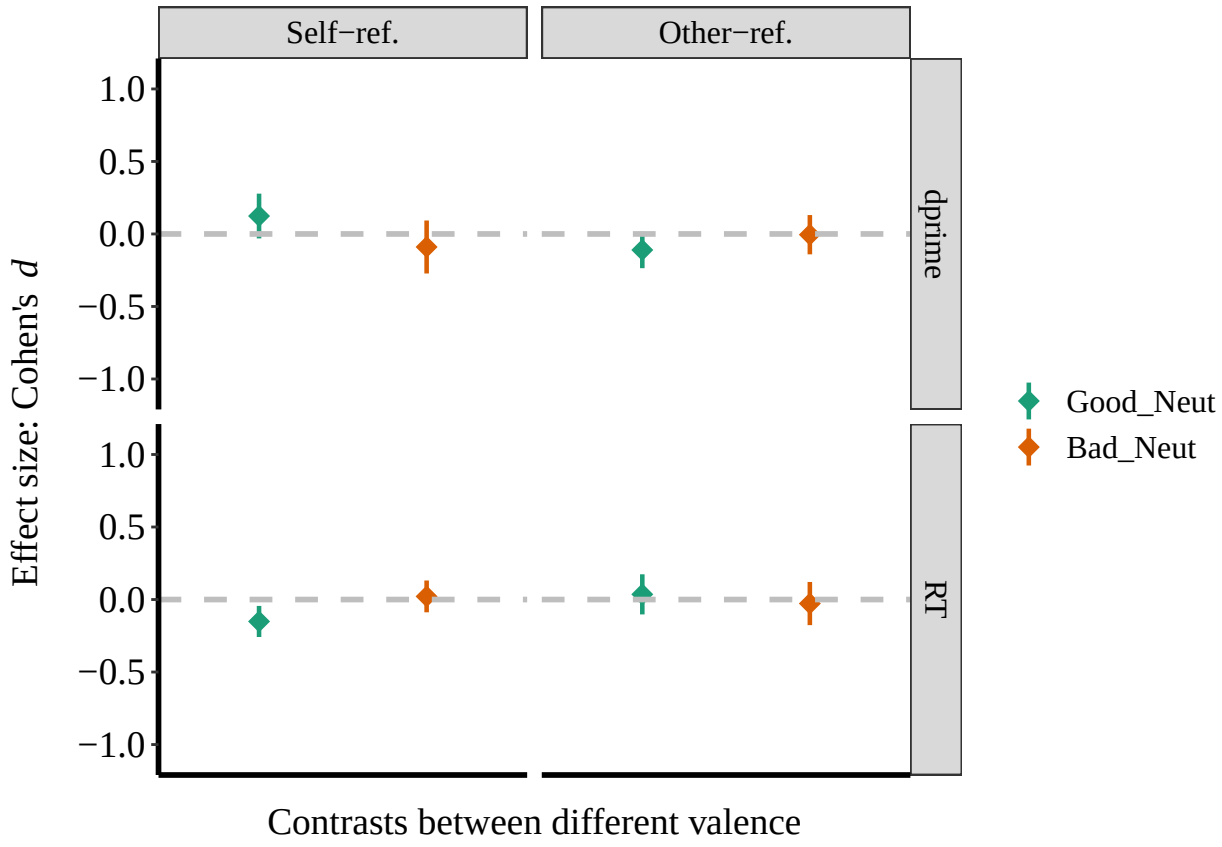

Figure 5. Effect size (Cohen's *d*) of Valence in Exp4a.

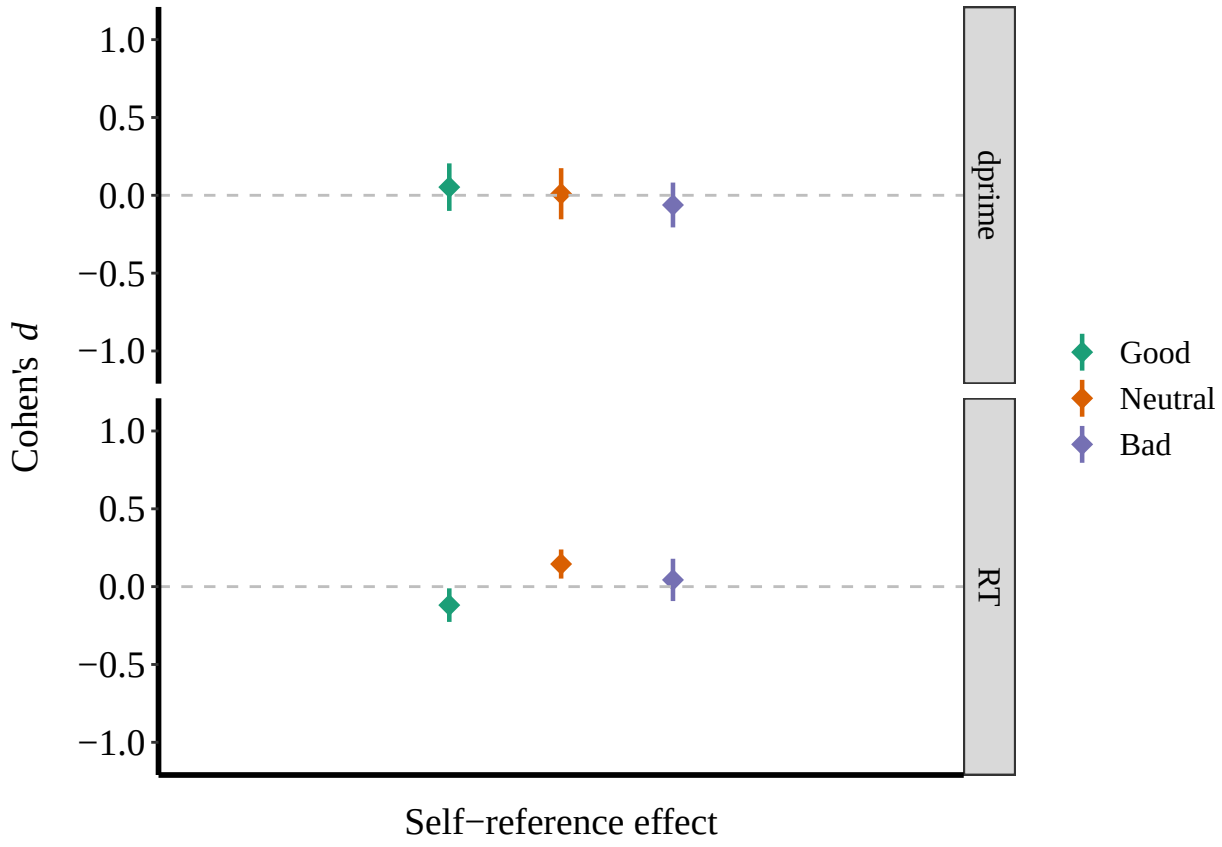

Figure 6. Effect size (Cohen's *d*) of Valence in Exp4b.

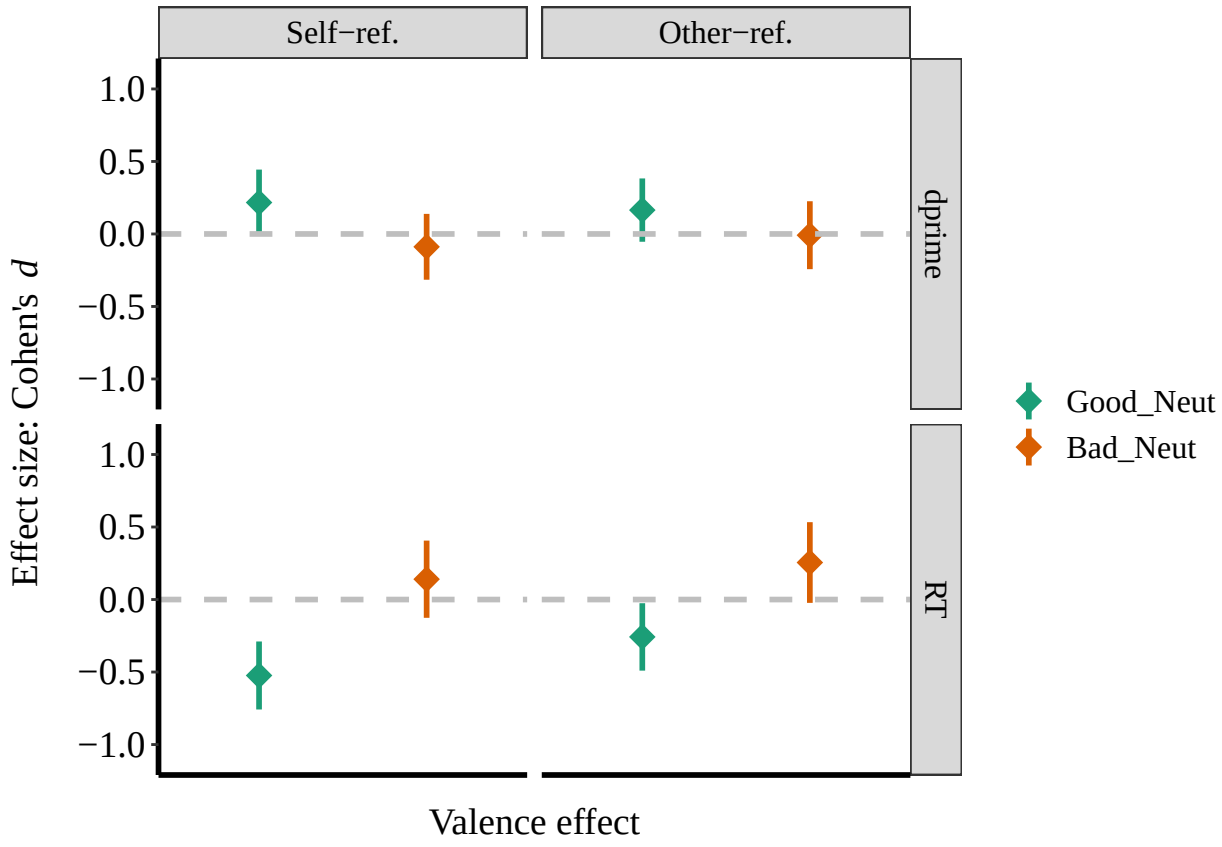

Figure 7. Effect size (Cohen's d) of Valence in Exp4b.

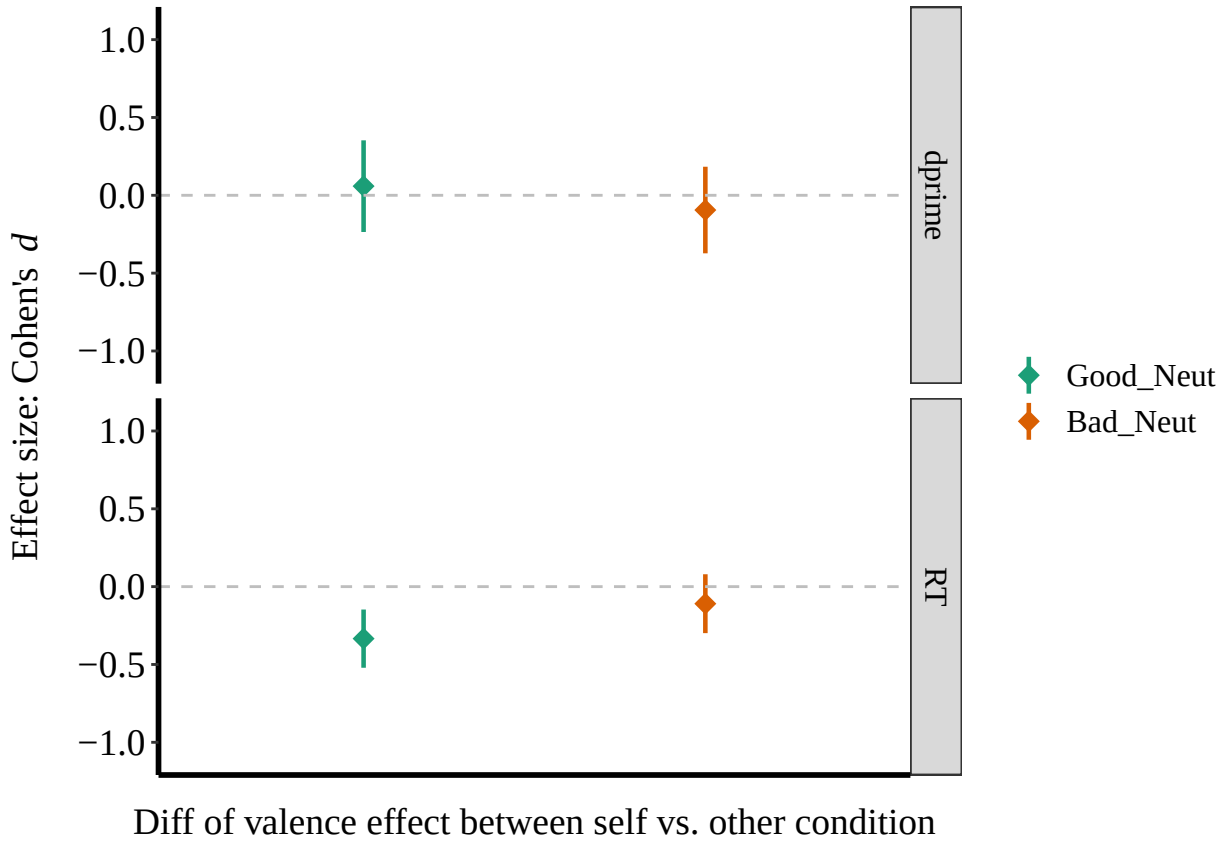

Figure 8. Effect size (Cohen's d) of Valence in Exp4b.

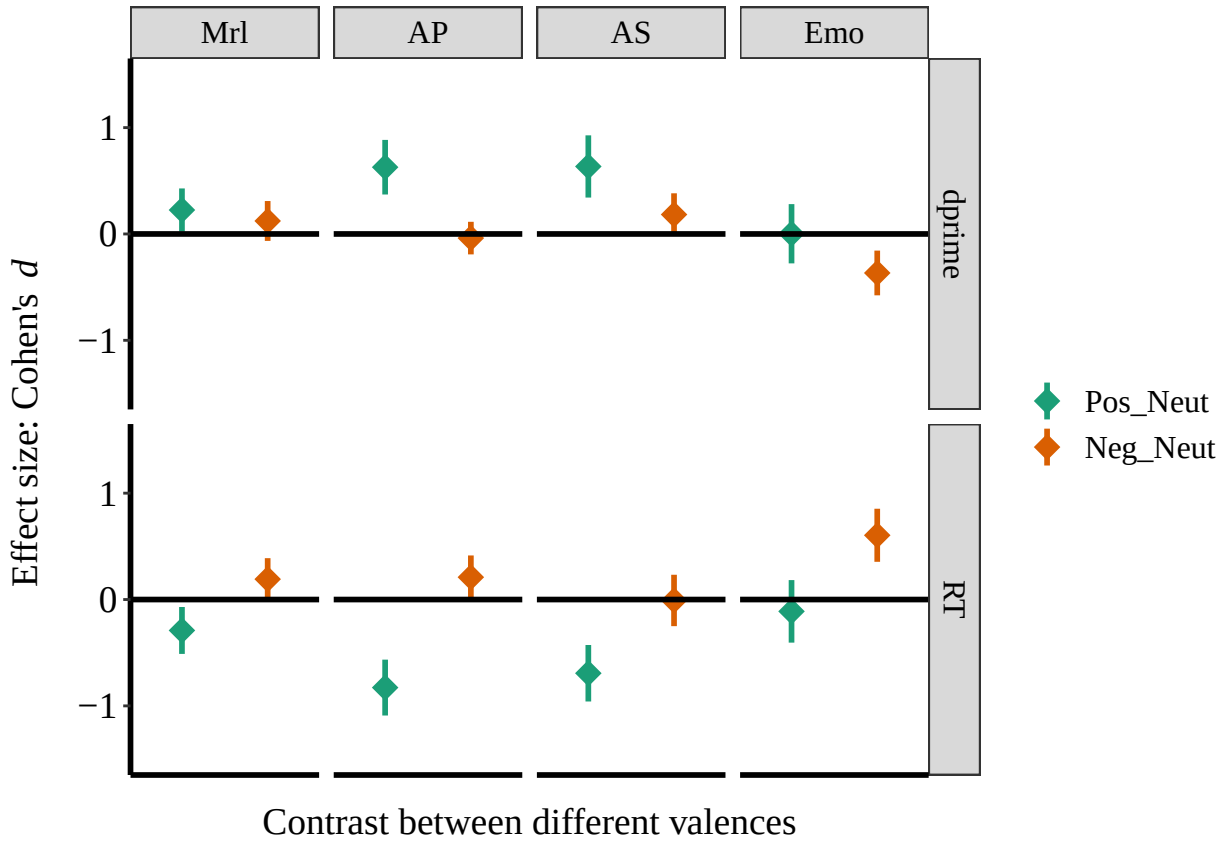

Figure 9. Effect size (Cohen's d) of Valence in Exp5.

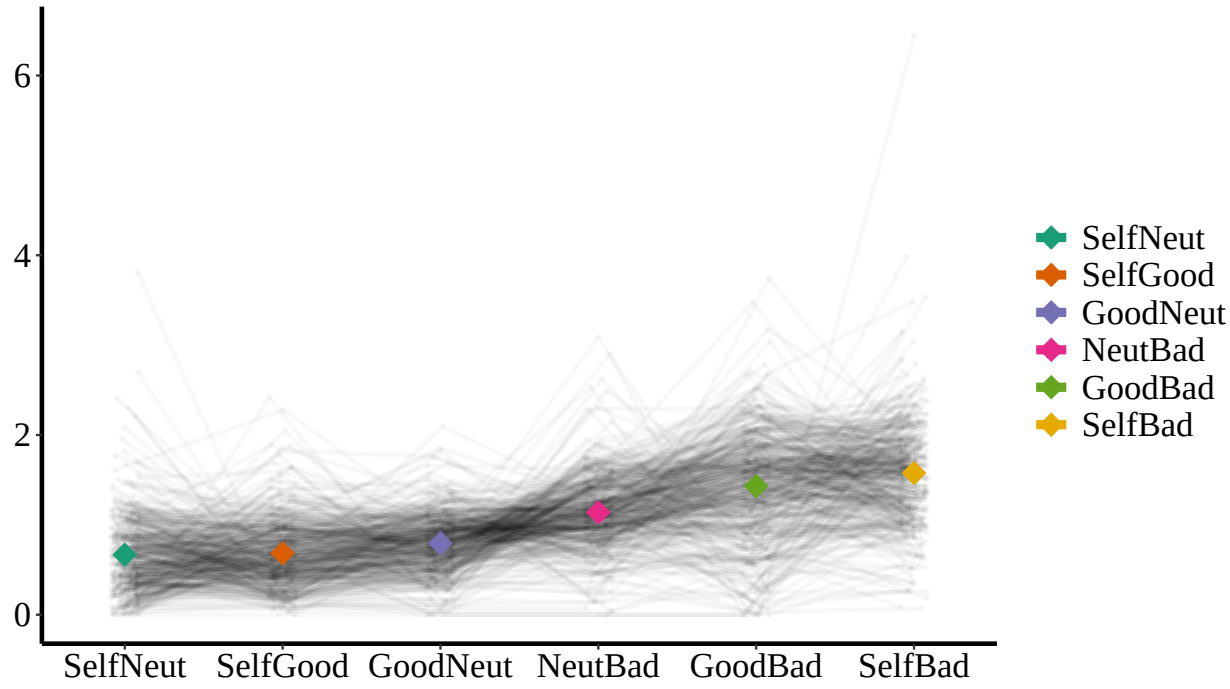

Figure 10. Self-rated personal distance

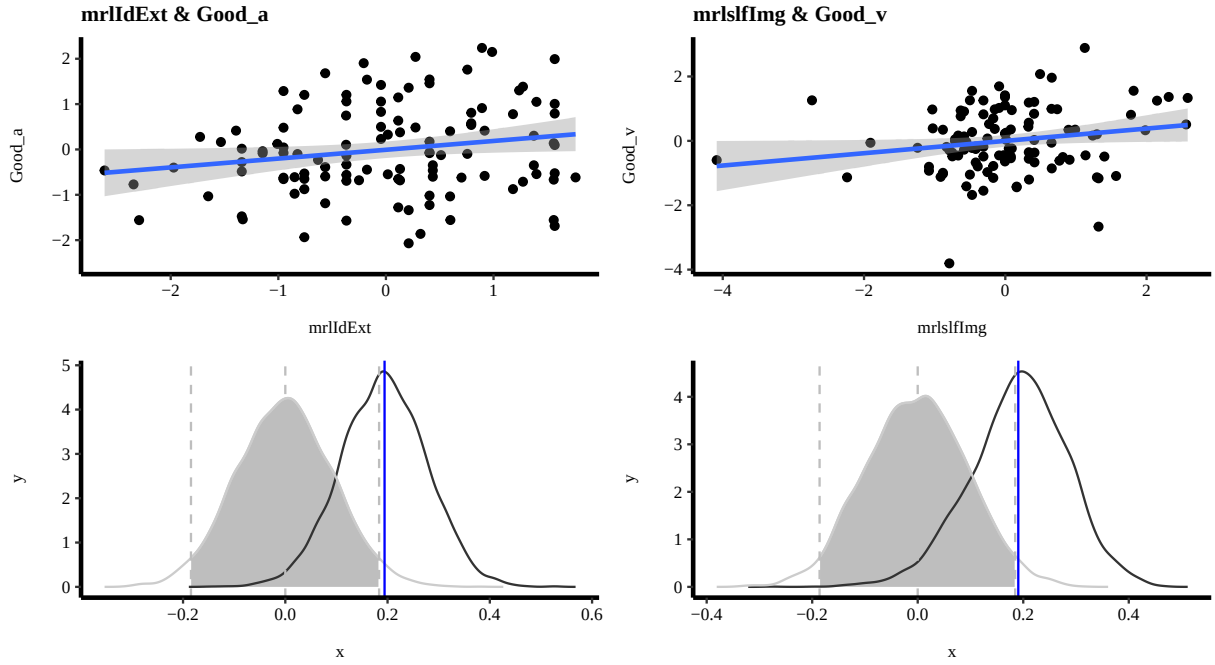

Figure 11. Correlation between moral identity and boundary separation of good condition; moral self-image and drift rate of good condition

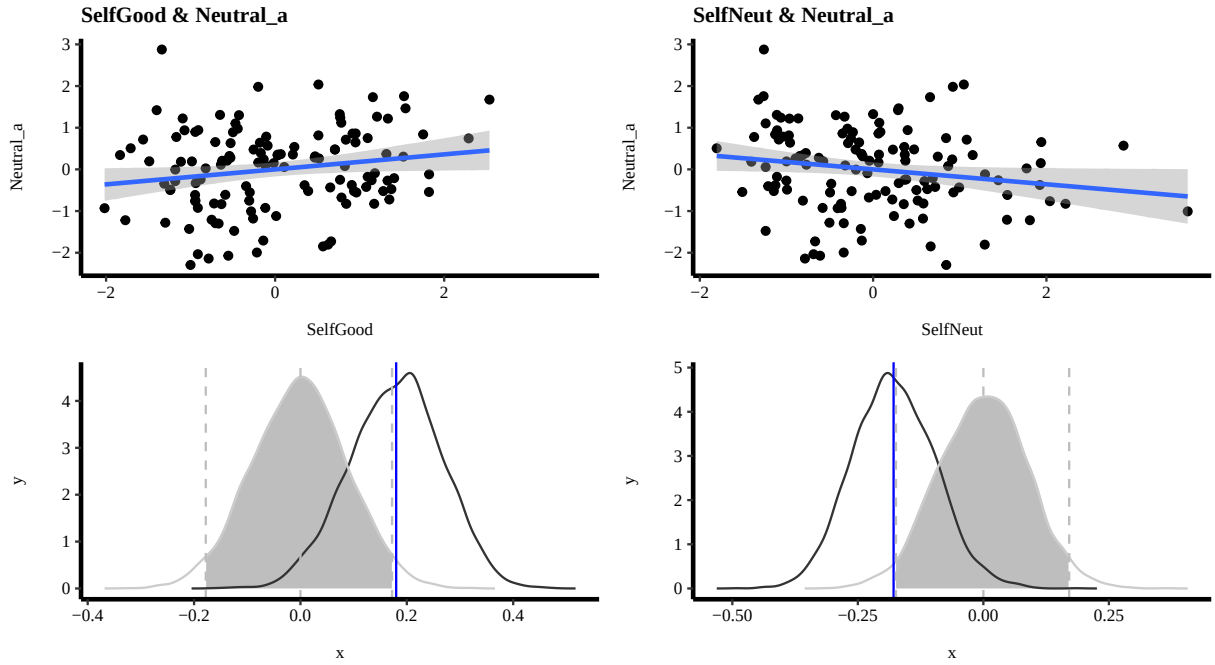

Figure 12. Correlation between personal distance and boundary separation of neutral condition

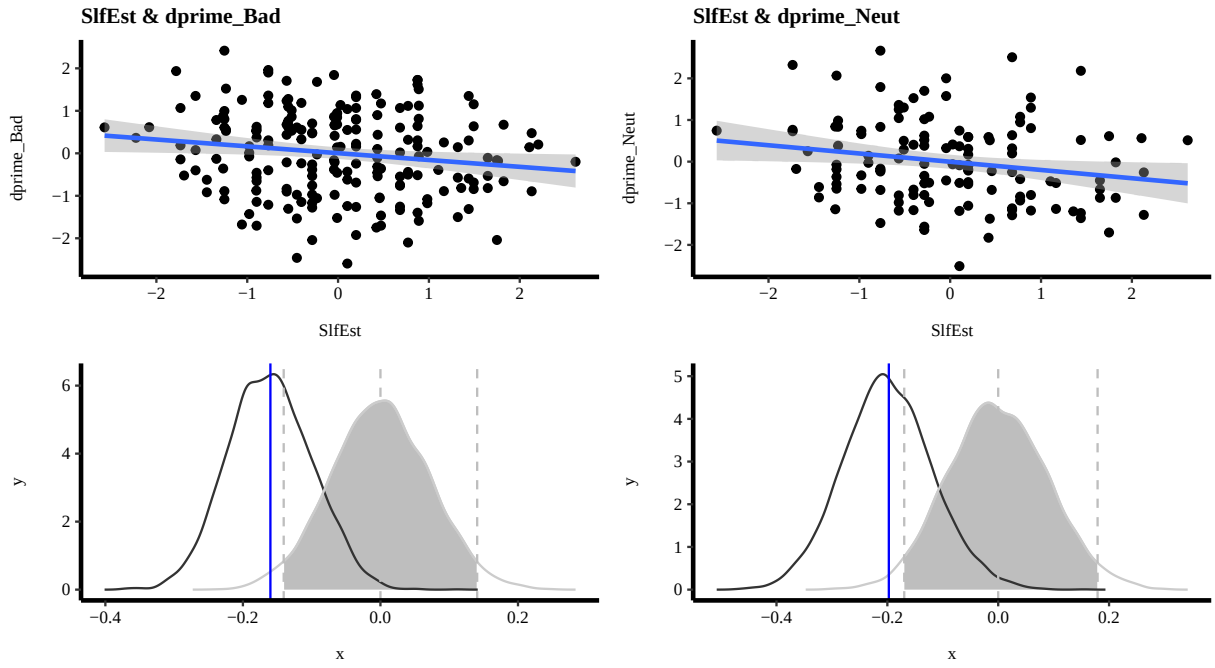

Figure 13. Correlation between self esteem and d prime of bad and neutral conditions