Title:
Differential valuation and learning from social and non-social cues in Borderline Personality Disorder

Authors:
Sarah K Fineberg MD PhD 1 *, Jacob Leavitt BA 2, Dylan S Stahl BA 3, Sharif Kronemer MS 4, Christopher D. Landry BA 5, Aaron Alexander-Bloch MD PhD 1, Laurence T Hunt PhD 6, Philip R Corlett PhD 1

1 Yale University Department of Psychiatry
2 University of Houston Department of Psychology
3 Yale University Child Study Center
4 Yale University Department of Neurology
5 Columbia University College of Physicians and Surgeons
6 Wellcome Centre for Integrative Neuroimaging, University of Oxford

* Corresponding Author. Address: Sarah Fineberg, Connecticut Mental Health Center Room 518, 34 Park Street, New Haven Connecticut, 06519. Telephone 203-974-7265. Email sarah.fineberg@yale.edu

Running Title: Learning in Borderline Personality Disorder

Keywords:
Borderline personality disorder
Associative learning
Prediction Error
Computational psychiatry
Social cognition
Trust

Word Count (abstract): 232
Word Count (paper): 3992
Number of references: 34
Number of figures: 5
Number of tables: 2
Supplemental information: methods and materials (347 words)

Abstract:

Introduction: Volatile interpersonal relationships are a core feature of Borderline Personality Disorder (BPD), and lead to devastating disruption of patients' personal and professional lives. Quantitative models of social decision making and learning hold promise for defining the underlying mechanisms of this problem.

Methods: In this study, women with BPD and healthy controls interacted with a confederate, then played an extended reward learning task that requires learning about non-social and social cue reward probability (The Social Valuation Task). Subject distress after debriefing was measured using a language marker (self-referentiality). Weighting of social and non-social cues in subject decisions was tested in mixed-effects regression models. Changes in learning rates in response to volatility for the non-social and social rewards was tested in a Rescorla-Wagner learning model.

Results: BPD but not control subject language had fewer distress markers after debriefing (including disclosure of deception). Compared to controls, BPD subjects weighted social cues more heavily, but they showed significantly blunted learning responses to reward volatility for both the non-social and social cues.

Conclusions: This is the first report of patient behavior in the Social Valuation Task. The results suggest that BPD subjects expect higher social and non-social volatility than do controls. These findings lay the groundwork for a neuro-computational dissection of social and non-social belief updating in BPD. Such an understanding, bridging circuity with clinical phenomenology through computation, portends clinical interventions that more directly target pathophysiology.

Introduction:

Learning whom to trust and when to revise trust attributions is a difficult but important task. People exhibiting extremes in trust can experience significant distress and personal risk, as in the very-low trust which characterizes paranoia (1) (2), and the very-high trust in Williams Syndrome (3) or amygdalar lesions (4). In Borderline Personality Disorder (BPD), trust is unstable, and interpersonal relationships are subject to recurrent episodes of rupture and repair. Low initial trust and rupture-promoting behavior in BPD have been modelled in the 10-round Trust Game, a brief economic exchange task played with a partner (5). We aimed to extend those data by examining responses to instability of social and non-social information.

Behrens et al. have examined associative learning about social cues, and, in so doing, they developed a measure of trust in a laboratory task setting (the Social Valuation Task (SVT)) (6). The SVT enables researchers to simultaneously examine non-social and social learning under uncertainty. In healthy undergraduates, learning rates increased with increasing reward volatility (7), and participants tracked non-social and social contingencies, weighing both to make decisions (6). FMRI identified prediction errors in canonical reward regions (striatum) for non-social information, and prediction errors in canonical social cognition regions (superior temporal sulcus / temporoparietal junction, dorsomedial prefrontal cortex, and middle temporal gyrus) for social.

Weighting of social versus non-social cues in the Social Valuation Task (SVT) in community samples is correlated with self-reported traits. Sevgi et al. screened a sample of healthy adults without psychiatric disorders for autistic traits using the Autism Quotient Spectrum Questionnaire (AQ), a self-report measure (8). People with low and high levels of autism traits in a modified SVT with implicit instead of explicit social cues (gaze direction instead of explicit advice). Autism score was directly correlated to poorer overall task performance (lower final point score), and inversely correlated to weighting social over non-social cues in the task. Subjects with high AQ scores were worse than low AQ subject at avoiding the influence of bad advice under volatile conditions (when the advice was most unreliable). In a similar design, Brazil and colleagues recruited 36 women from the community with a range from low to high score on the Psychopathic Personality Inventory (a 187 item self-report measure of psychopathic traits in non-clinical individuals) to play the SVT in

the lab (9). As with autism score, one of the psychopathy sub-scales (social potency – a measure of the ability to charm and manipulate others) was inversely correlated with weighting of social data. The stress immunity sub-scale (a measure of having low anxiety) was inversely correlated with use of either non-social or social cues to make decisions in the SVT. The fearlessness subscale was specifically inversely correlated with use of the non-social cue.

We report here the first test of SVT behavior in a patient population. We model both weighting of social versus non-social cues and learning rates in response to changes in social and non-social reward volatility.

We had several main hypotheses:

People with BPD would be highly sensitive to social cues:

1) People with BPD would more heavily weight social than non-social cues.
2) People with BPD would weight social cues more than did control subjects.
3) People with BPD would weight negative social cues more and positive social cues less than did control subjects.

People with BPD would be sensitive to social reward volatility:

4) People with BPD would increase learning rate more than did controls with increasing social volatility.
5) People with BPD would increase learning rate more in response to social volatility than in response to increased non-social reward volatility.

Methods and Materials:

Ethics

This protocol was written in accordance with the Declaration of Helsinki and approved by the Yale Institutional Review Board (HIC protocol 1211011104). All activities were conducted according to the protocol.

Subjects

Women aged 18-65 were recruited from the community. Screening for psychiatric symptoms was conducted by phone screen using the Diagnostic Interview for Personality Disorders BPD questions (10). People likely to either have no psychiatric diagnosis or have BPD were then invited for in-person semi-structured interview with a psychiatrist (SKF) including the Revised Diagnostic Interview for Borderlines (DIB-R (11)) and Structured Clinical Interview for DSM-IV (SCID (12)). To be included in the study, controls had no psychiatric conditions, and BPD subjects had no current substance dependence and no primary psychotic disorder. On the DIB-R, controls scored ≤ 4 (scaled total), BPD subjects scored ≥ 8 (scaled total)). Included subjects also read English well (no history of special education, ≤ 11 errors on the Wide Range Achievement Test 4th Edition (WRAT-4) reading test (13)), had intact color vision, had no history of head injury or neurologic condition, and had no gait disturbance. See demographics in **Table 1**.

We also recorded subject education level, hours of work and/or schoolwork per week, current relationship status, and reading level (to detect differences among the more literate subjects in our sample, we used the more challenging North American Adult Reading Test (NAART) instead of the WRAT-4 score here) (14). See **Table 2**.

Self-report scales

The Borderline Symptom List (BSL-23), Beck Anxiety Inventory (BAI), Beck Depression Inventory (BDI), and the Structured Clinical Interview for DSM-IV Personality Questionnaire. See details about these scales in the Supplementary Methods and Materials.

Social Valuation Task Design

On each of 290 trials (~45 minutes), subjects were presented with a binary choice between a green box and a blue box (non-social cue) **(Figure 1A)**. They also saw a point value associated with each color, and advice from a confederate about which color to pick (social cue). Before play, the subject met the confederate, interacted briefly, and together, subject and confederate were instructed in how to play the game. They were instructed to attend to three main concerns: how to choose a color, how to use the advice, and how points work. Specifically that: 1) one color would be more likely to be correct, and through feedback after each trial, the subject would be able to learn which color to

choose, 2) that the confederate would give advice to the subject that may be helpful, or that may serve the confederate's own goals, and 3) that points are displayed on each of the two choices (green and blue) for each round, with correct choices rewarded by the displayed point value, and no possibility of losing points. Correct choices add the displayed point value to a total point tally. The tally was displayed as a growing bar at the bottom of each screen. If the bar grew long enough, the subject would win a $10 or $20 bonus. All subjects were in fact compensated $20 for the study visit and task completion. During a demonstration round, the subject observed the confederate choose to "give correct advice" and "give incorrect advice".

The task includes orthogonal variation in non-social (color) and social (advice) reward probability and reward volatility (frequency of changes in reward probability) (**Figure 1 B-D**). Behrens et al. found that healthy control subjects learn from variation in both non-social and social reward contingencies in the SVT (6).

Confederate

The task confederates were all 20 - 30 year old Caucasian women trained for consistent interaction with subjects and consistent performance during the demonstration task.

Debriefing

Immediately after the end of the task, subjects were audio-recorded in a discussion with study staff in response to a list of specific questions and statements about the task experience. We asked 4 questions before the disclosure that the confederate was not actually a second game player, then 2 more questions after the disclosure. We examined the transcribed language from the debriefings. We counted the number of times that the subject mentioned the confederate.

To capture shifts in emotional state before versus after the disclosure, we examined the use of self-referential pronouns in subject speech (15). Transcribed speech was analyzed with Linguistic Inquiry and Word Count (LIWC) (16), which

returns frequency of specific categories (we used "first-person pronouns") as count/total words. We used repeated-measures ANOVA to test for interaction between time (before versus after disclosure) and group (BPD versus control).

Modelling task behavior: relative cue weighting

Variables influencing subject decisions were examined with mixed models in the statistical program R using the package 'lme4' (Bates et al. 2014). Probability and volatility values were those derived by Behrens et al. from their Bayes-optimal model (7). Non-social variables were points (difference between point magnitude for green and point magnitude for blue), probability green is correct, and volatility of green being correct. Social variables were current trial advice, current advice weighted by probability advice is correct, current advice weighted by volatility of advice being correct, and refusing current advice after recent betrayal. We also tested a series of time windows on recent betrayal (incorrect advice) or help (correct advice): within x trials, where x = 1, 3, 4, 5, 6, or 7. Each variable was centered and z-scored to facilitate comparison of coefficients across factors.

The impact of clinical group was tested separately for each of the above predictor variables, *v* (modelled as fixed effects in the mixed models). All models also included subject as a random variable (modelling both random intercept and random slope). Four models were compared for each variable: a full model including both clinical group and the interaction *v*-by-group as fixed effects (model #); a reduced model without the interaction effect (model p); a further-reduced model that included neither the interaction nor the main effect of clinical group (model n), and a fully-reduced model that contained only the random-effect for subject (model x). *P*-values were obtained by likelihood ratio (chi-square) tests for model # compared to model p (to test for a significant interaction effect), for model p compared to model n (to test for a significant effect of clinical group), and for model n compared to model x (to test for a significant effect of predictor).

In order to test for apparent statistical differences that may be explained by our having treated each binary decision as independent in the statistical model, we verified the results with a permutation procedure where group membership is

randomized and test statistic recalculated. We found statistically significant ($p < 0.05$) values for likelihood ratios (chi-square) tests comparing the permuted to the reported models.

Modelling task behavior: learning rates

Subject learning rates were modelled using the R package "hBayesDM" and the function "bandit2arm_delta" using default parameters (http://rstudio-pubs-static.s3.amazonaws.com/164729_b8c91397d2d648cd928734286053e905.html#list-of-tasks-and-computational-models-implemented-in-hbayesdm) (17). This package calculates a mean learning rate for each subject based on the Rescorla-Wagner (delta) equation. See details in the Supplementary Methods and Materials.

In our analysis, learning from the non-social cue in the SVT was modelled based on the trials in each of 2 task phases: stable (trials 1 – 130) and volatile (trials 131 – 290). Learning from the social cue was modelled based on the trials in each of 3 task phases: stable reliable (trials 1-70), volatile (trials 71 – 210), and stable unreliable (trials 211 – 290). Repeated-measures ANOVA was used to test for group x phase interaction for each cue.

Results:

Demographics

Subjects did not differ significantly in age (t-test t statistic = -0.47, df = 39, p = 0.641) or education (t-test for years in school, t-statistic = 0.751, df = 39, p = 0.457, for reading score p = 0.53) (see **Table 1**). BPD subjects were significantly more symptomatic than controls with respect to self-reported BPD symptoms, depression symptoms, and anxiety symptoms (**Table 2**). All the subjects were able to complete the task, and their final point scores did not differ by group or symptom burden (**Figure 2A**).

BPD patients show lower implicit distress in response to task

As a preliminary test for the expected focus on social cues in BPD versus control, we counted references to the confederate in the audio recordings of the post-task debriefing. Subjects with BPD talked much more about the confederate than did the control subjects (**Figure 2B**). There were no between group differences in our subjective ratings of surprise (mean BPD = 0.8, SE = 0.29; mean control = 0.77, SE = 0.23; t-test t statistic= -0.08, df = 21, p = 0.93), distress (mean BPD = 0, SE = 0; mean control = 0.15, SE = 0.10, t-test t statistic = 1.3, df = 21, p = 0.21), or suspicion (mean BPD = 0.10, SE = 0.10; mean control = 0.08, SE = 0.08, t-test t statistic = -0.19, df = 21, p = 0.85).

Though none of the subjects demonstrated overt distress during or after the task, we expected that they may have experienced implicit distress due the periods of volatile and untrustworthy advice in the task, or after hearing at the end that the confederate was not in fact another player, and that we had intentionally misled them. As a measure of implicit distress, we measured self-referential language (words like I, me, and mine), which has been demonstrated to increase with depression, suicidality, and other mental and physical illnesses (15, 18) (19). Control subjects did not change their use of self-referential language rom before to after the post-task disclosure, but BPD subjects used significantly fewer self-referential words after the disclosure (**Figure 2C,** time x group interaction F = 6.16, p = 0.02), suggesting that their distress actually decreased after it was revealed that they were interacting with a computer, rather than a human confederate.

H1/2: People with BPD weighted social more heavily than non-social cues

We tested the impact of non-social and social cues on subject choices in the SVT (**Figure 3**). To test our first and second hypotheses, we examined the weights of non-social and social cues in subject decisions. Each of the examined variables was a significant contributor to subject decisions, and contributed differently to decisions between groups. Specifically, BPD participants were more likely to choose green when reward probability was higher (**3A,** chi-square statistic from likelihood ratio test = 4.03, p =0.045) and a trend towards being less likely to choose green when likelihood of reward became more volatile compared to controls (**3C** chi-square statistic from likelihood ratio test = 3.48, p = 0.062). They were also more likely to choose green if the difference between points for green and points for blue was larger (**3E**, chi-square statistic from likelihood ratio test = 4.07, p < 0.044) than controls. BPD participants were more likely to go with

the advice if reward probability was higher compared to controls (**3B,** chi-square statistic from likelihood ratio test = 5.98, p = 0.015).

Of interest, and perhaps surprising, is that both groups (BPD > control) were also more likely to take the advice when social reward likelihood was more volatile (**3D,** chi-square statistic from likelihood ratio test = 4.96, p < 0.026). However, the model describing outcomes predicted by group and current trial advice (what Behrens et al. termed "blindly following advice") did not detect statistical differences (**3F**). In sum, we found that people with BPD made significant use of both non-social and social cues. Interestingly, people with BPD weighted both social **and non-social** cues more heavily than controls, though, between-group differences were larger for weighting of social reward probability than for non-social reward probability (based on magnitude of difference between regression lines; as noted chi-square tests were significant in both cases).

H3: People with BPD used positive social cues more than negative social cues to make decisions

To better understand the responses of BPD and control subjects to social cues in this interactive context, we next tested the predicted choices after recent betrayal (bad advice) or help (good advice) (**Figure 4**). We found a significant decrease in BPD patients in use of betrayal to avoid bad choices (**4A,** BPD> control, for betrayal within the last three trials, chi-square statistic from likelihood ratio test = 4.25, p= 0.039) and a trend toward a group x predictor interaction for increased use of help to find good choices in BPD patients (**4B, for help within the last 3 trials,** BPD > control, chi-square statistic from likelihood ratio test for group x predictor interaction = 3.79, p = 0.051).

We also tested the rate of decrement in weighting of recent help or betrayal. As expected, both groups used help or betrayal less as the window size expanded (**Figure 4C,D**), but both help and betrayal were significant predictors of outcome out to at least a 7-trial window. However, it was help (the positive social cue) not betrayal (the negative social cue) that showed a trend towards a group x predictor interaction (use of help decayed more slowly in the BPD than the control group).

H4/5: Learning rates reveal blunted response to increased reward volatility in people with BPD

We modelled learning rates for non-social and social rewards during the stable and volatile phases of the task. We found that control and BPD subjects learned at similar low rates about non-social data during the initial phase when reward probability was stable (**Figure 5A,** $t = -1.38$, $p > 0.05$). However, when reward probability became volatile (phase 2), control subjects increased their learning rate more than twice as much as did the BPD subjects (**Figure 5A**, significant group x condition interaction: $F = 19.78$, $p < 0.001$). Learning from social cues was slower in BPD than in controls during all three phases of the task **(Figure 5B,** stable reliable $t = 4.02$, $p < 0.001$, volatile $t = 2.90$, $p < 0.01$, stable misleading $t = 3.44$, $p < 0.005$), and response to volatility in BPD was significantly lower than in controls (**Figure 5B,** group x condition interaction, $F = 5.81$, $p < 0.01$). These results were surprising: we had hypothesized faster learning rates in BPD in response to reward volatility, and instead, we observed blunted response compared to controls for both non-social and social cues.

Discussion:

In this first study of the SVT in a patient population, we examined task performance in people with BPD, a condition defined by prominent interpersonal problems. In this extended interactive paradigm, women with BPD did indeed focus on social experience, weighting social over non-social cues to make decisions. However, we also found that a negative social experience (incorrect advice) was a less potent and less durable influence on subject choice than a positive social experience (correct advice).

Previous work in non-interactive paradigms, such as Reading the Mind in the Eyes and morphed face challenges has identified a strong negative attribution bias (reviewed in (20, 21)). BPD patients attend quickly to negative faces and spend more time looking at them. A small number of studies have tested the response of BPD patients to social interaction games using brief paradigms. In the 10-round "Trust Game", players with BPD responded with low initial trust and failure to coax defecting partners back to play (5).

The hormone oxytocin is often associated with social bonding (22), but also thought to amplify social behavior of both positive and negative valence depending on context. Bartz et al. found that after oxytocin dosing, women without BPD were more trusting and cooperative in an economic exchange game. By contrast, women with BPD were less trusting and cooperative, suggesting differential responses to this neurochemical social signal.

In our sample of women with BPD and controls, we found that both non-social and social cues were weighted more heavily in the BPD group than in the control group. This may suggest that people with BPD are more attentive to all of the cues around them. Previous work describing learning in BPD has had mixed results. Work in brief non-social paradigms found that BPD state (emotional arousal) but not traits (BSL score) predicted problems with learning acquisition and vice versa for reversal learning (23). Others found no difference in reversal learning (24), but deficits in working memory in BPD (25).

In the extended and more complex social interaction in the SVT, we were able to examine not only low-probability but also low-reliability social rewards. In contrast to the defection (failure to coax) that was observed in the trust game after low-payoff trials (5), we saw increased use of social cue under conditions of high social reward volatility here (in both groups, but BPD > control), as if subjects were redoubling their efforts to remain socially engaged.

Unlike the reported problems in task performance with increased sub-clinical autism and psychopathy symptoms, our sample of women with BPD completed the SVT with final point scores similar to controls, and used both social and non-social cues to make decisions.

We extended previous reports of the effect of personality/mental health traits on SVT behavior by examining learning rates. We replicated Behrens et al. report that control subject learning rates increase under conditions of reward volatility. However, we were surprised to observe that BPD subjects showed only half the learning rate increase that control subjects did in response to non-social reward volatility, and barely responded at all to social reward volatility. One possibility is that the BPD subjects assume higher baseline volatility of all environments and contingencies, such

that high volatility is not surprising, and does not prompt updating. This is consistent with research demonstrating early life adversity, especially neglect (= volatility) is a key risk factor for BPD. Our observation that people with BPD decrease their self-referential language after confirmation of deception is consistent with this idea. The BPD subjects used fewer language markers that connote distress once they were informed of the deception, consistent with the idea that they harbor assumptions that the social world is unreliable. For someone sensitive to volatility, attending closely to cues, but not updating rapidly may be adaptive.

This insensitive learning account of BPD social behavior is also consistent with a recent report from Read Montague's group describing a computational model of BPD trustees' responses in the 10-round Trust Game (26). They had previously found that people with BPD fail to coax a defecting partner to re-engage in economic exchange. They now report that a hierarchical belief model significantly benefits from parameters describing one's own irritability and beliefs about the partner's irritability. Here, irritability means the likelihood of retaliation on a low economic offer. BPD players were significantly less sensitive to their partners' irritability than control players, and the authors suggest that this leads to missed opportunities to respond: a player doesn't coax if she doesn't detect early cues that the partner is becoming irritable or is likely to disengage.

Particular strengths of this work include the use of a patient population, in fact, we carefully screened participants for non-clinical status (controls) or significant symptoms (DIB score > 8 in the BPD group). Our subjects met and interacted with the confederate before starting the task, perhaps increasing their ability to engage with the task in a manner that reflects their real-world social behavior. The Social Valuation Task is a lengthy interactive task that combines non-social (one's own beliefs) with social (others' counsel) for decision-making at each trial. The task architecture includes orthogonal periods of volatility for the non-social and social cues which allows us to model the relative use of each data source in decision-making.

Limitations of this work include the relatively small sample size and the focus on only female subjects (though this was important given the small sample size, we cannot generalize these results to men at this point). Though we mostly see

our very symptomatic patient group as a strength of this study, our exclusion of people with fewer or less intense symptoms does preclude dimensional analysis of the impact of symptom burden on behavior here.

Our sample was mixed in terms of their experience with treatment, and our group size was too small for subgroup analyses, but we would be very interested to investigate in future studies the impact of past or current treatment on SVT performance, and how behavior in laboratory learning tasks may help predict response to psychotherapy.

In future work, we aim to test social learning in BPD in settings that even better engage real-world relationships, to take a dimensional approach to relating learning to specific symptoms and treatment response, and to query the neurobiology underlying the observed differences.

Acknowledgements:

We would like to thank our research participants. Several members of our lab helped as confederates and with recruitment efforts: Emily Finn, Carol Gianessi, Kristen Budde, Erin Feeney, Margot Reed, Sasha Deutsch-Link, Megan Ichinose, Taylor McGuiness, and Rachel Zubi. Clinicians at the Yale New Haven Hospital Intensive Outpatient Program and in the Connecticut Mental Health Center Ambulatory Services also helped with subject recruitment. Megan Ichinose, Lindsey Conkey, Mary Zanarini, and Emily Ansell gave very helpful advice on research methods.

This work was supported by the Connecticut State Department of Mental Health and Addiction Services through the Connecticut Mental Health Center Clinical Neuroscience Research Unit. Sarah K. Fineberg was supported by NIMH Grant no. 5T32MH019961, "Clinical Neuroscience Research Training in Psychiatry" and a NARSAD Young Investigator Award (2014-2016). Dylan S. Stahl was supported by the Richter Memorial Fund and the NSF COAST Award as an undergraduate student at Knox College. Aaron Alexander-Bloch was supported by was supported by NIMH Grant no. 5T32MH019961, "Clinical Neuroscience Research Training in Psychiatry". Laurence Hunt was supported by a Sir Henry Wellcome Fellowship from the Wellcome Trust (098830/Z/12/Z), and a NARSAD Young Investigator Award. Philip R. Corlett was funded by an IMHRO/Janssen Rising Star Translational Research Award and CTSA Grant Number UL1TR000142 from the National Center for Research Resources (NCRR) and the National Center for Advancing Translational Science (NCATS), components of the National Institutes of Health (NIH), and NIH roadmap for Medical Research. The contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

Other results from other experiments conducted in this patient sample have been published elsewhere (https://www.ncbi.nlm.nih.gov/pubmed/29248760). Preliminary versions of these results have been shown as a poster and an oral presentation at the National Association for the Study of Personality Disorders and as a poster at the Society of Biological Psychiatry.

The authors report no conflicts of interest or financial disclosures.

	Control	BPD
n	23	20
Age (yrs)	33.86 +/- 2.93 (18-60)	35.80 +/- 2.91 (18-63)
Education (yrs)	14.78 +/- 0.57 (10-19)	14.2 +/- 0.54 (11-20)
Ethnicity		
Asian	13%	10%
Black	30%	15%
Hispanic	9%	5%
White	44%	55%
Not reported	4%	15%
Taking psychiatric meds	0	45%
Anti-depressant	0	15%
Mood Stabilizer	0	25%
Anti-psychotic	0	15%
Benzodiazepine	0	10%
Current relationship		
None	30%	35%
In a relationship	26%	35%
Married	13%	10%
No answer	31%	20%
Has children?		
Yes	22%	17%
No	48%	55%
No answer	30%	28%
Current work		
None	26%	45%
0-20 h/wk	13%	15%
20+ h/wk	22%	15%
In School	30%	10%
No answer	9%	15%

Table 1. Subject demographics. Note that some individuals in the BPD group were taking multiple psychiatric medications.

	Control	BPD	p
NAART	20.90 +/- 2.09	19.20 +/- 1.69	0.53
SCID2-BPD	0.90 +/- 0.28	9.00 +/- 0.76	<0.001
BSL	5.95 +/- 1.60	33.00 +/-4.29	<0.001
BDI	3.10 +/- 1.06	21.40 +/- 2.95	<0.001
BAI	7.75 +/- 2.32	23.00 +/- 2.86	<0.001

Table 2. Subject characteristics. Mean and standard error of the mean are displayed for the NAART (reading test), two different BPD self-reports (SCID2-BPD and BSL), depression self-report (BDI), and anxiety self-report (BAI). T-tests comparing control to BPD subject scores revealed no difference in reading score between groups, but significant differences in each of the self-reports.

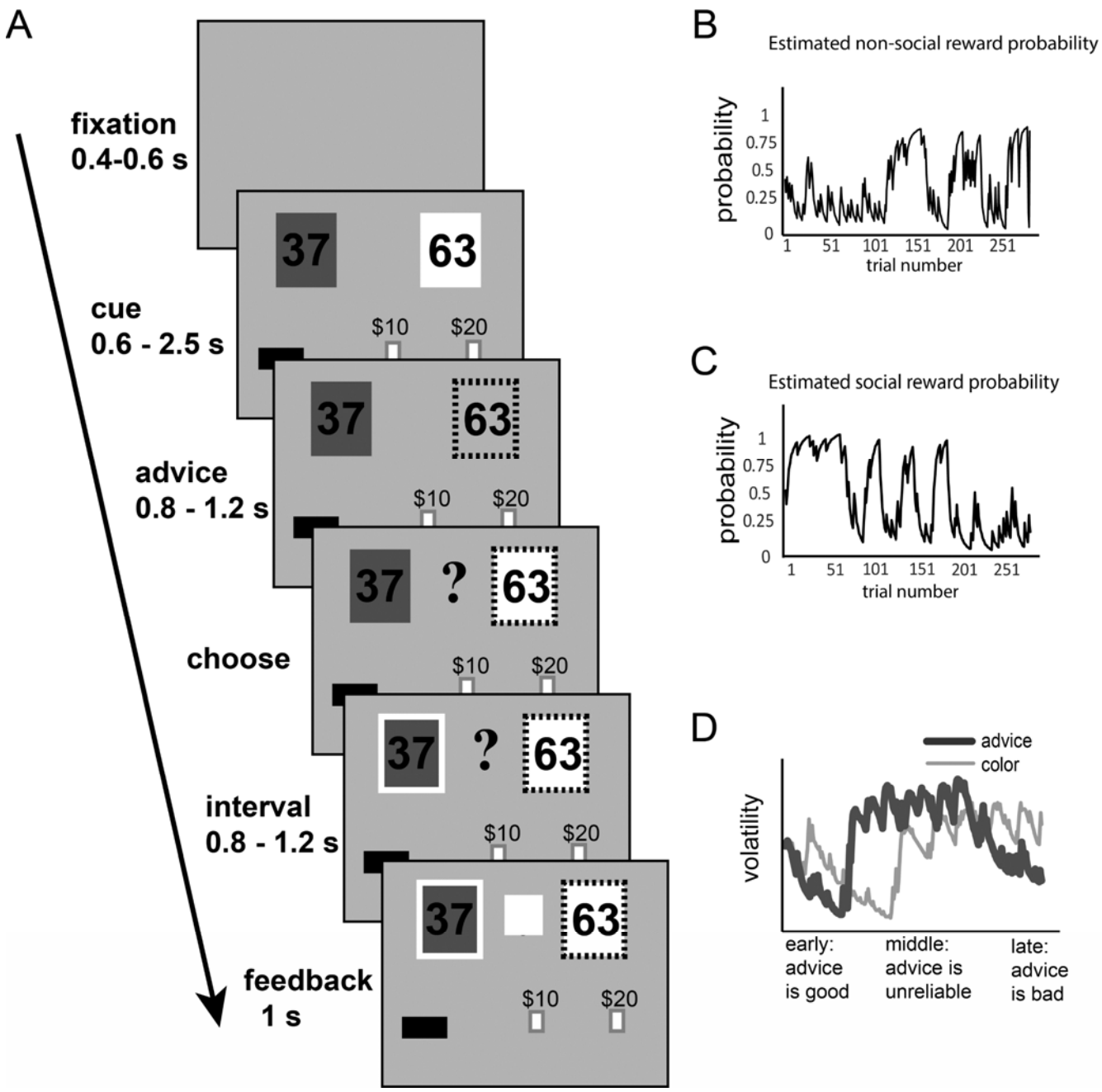

Figure 1. Task Design A) Cartoon of task phases including elements displayed to the subject in each phase, and timings. Note that the subject makes her choice in the "choose" phase, and sees the correct answer in the "feedback phase". These six phases are repeated in each of 290 trials. B) Changing probability of reward for the non-social cue (green) over time. C) Changing probability of reward for the social cue (advice) over time – note separate pattern from non-social cue. D) Changes in the volatility of the non-social (thin line) and social (thick line) reward probabilities represented as volatility over time. X-axis is again time (trial number), and quality of advice is indicated as it changes from low-volatility and trustworthy advice at the beginning through a period of high-volatility advice, to a period of low-volatility but untrustworthy advice at the end of the task.

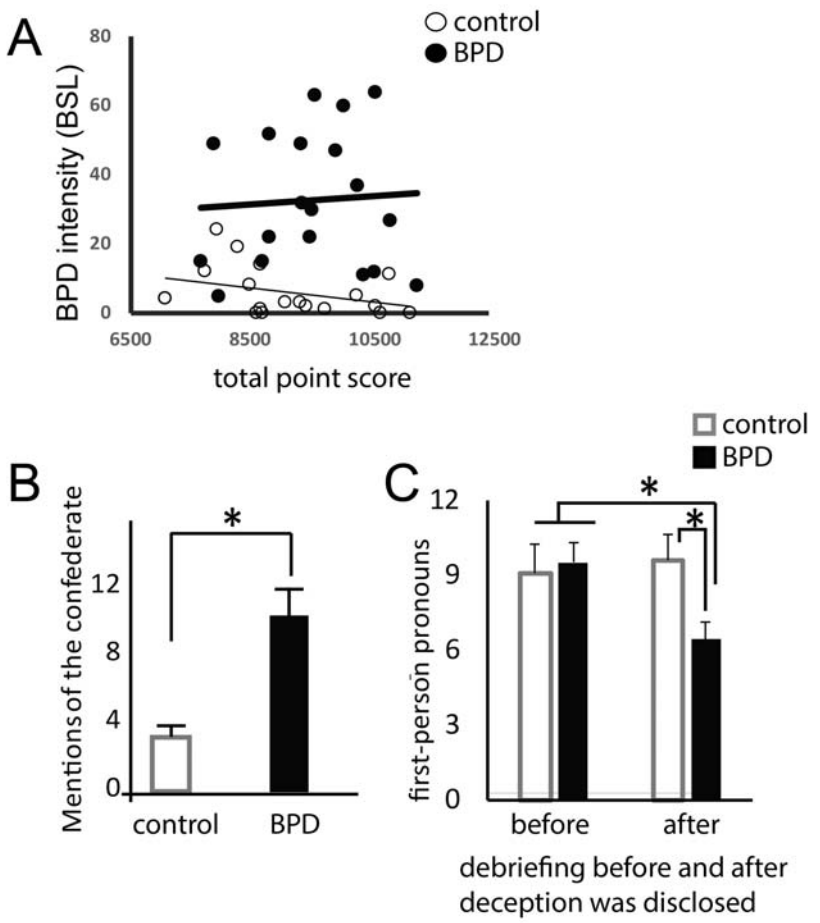

Figure 2 Task Experience A) BPD and control subjects achieve similar final point scores in the game (t-test p = 0.345), and final point score does not correlate with BPD symptom score in either group (Pearson correlation Control r = -0.352, p = 0.140, BPD r = 0.061, p = 0.804). **B)** In the post-task debriefing, BPD subjects talk significantly more about the confederate than do control subjects. Error bars represent standard error of the mean (t-test p = 0.01). **C)** In the post-task debriefing, BPD and control subjects refer to themselves with similar frequency before the deception is revealed. However, after they hear that the computer, not the confederate, was providing the advice, self-referential language is significantly less in BPD than in control subjects. In a repeated-measures ANOVA, there was a trend-level difference by time (F = 3.03, p = 0.09) and a significant difference for the time x group interaction (F = 6.16, p = 0.02).

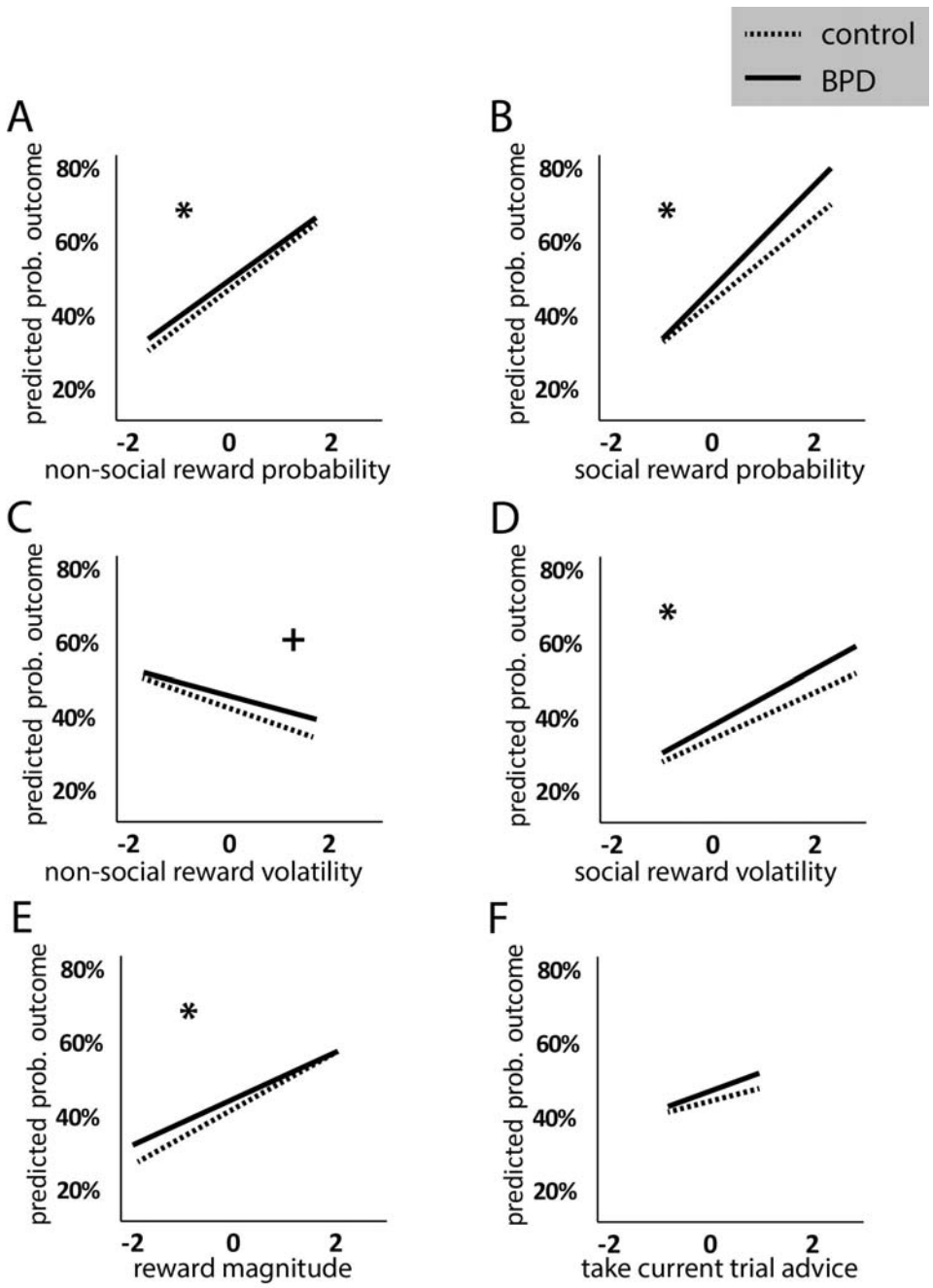

Figure 3. BPD and control subjects use cues differently to make decisions
Mixed-effect logistic regression models were run to measure the effect of non-social and social predictors on subject choices. Predicted probability of choosing green is plotted over z-scored values of each predictor. Non-social predictors (A,C,E) and social predictors (B, D) both performed differently between groups (except F, current trial advice). Significant Chi-square for models with versus without group term at $p < 0.05$ is indicated by *. Trending Chi-square for models with versus without group term at $p < 0.1$ is indicated by +.

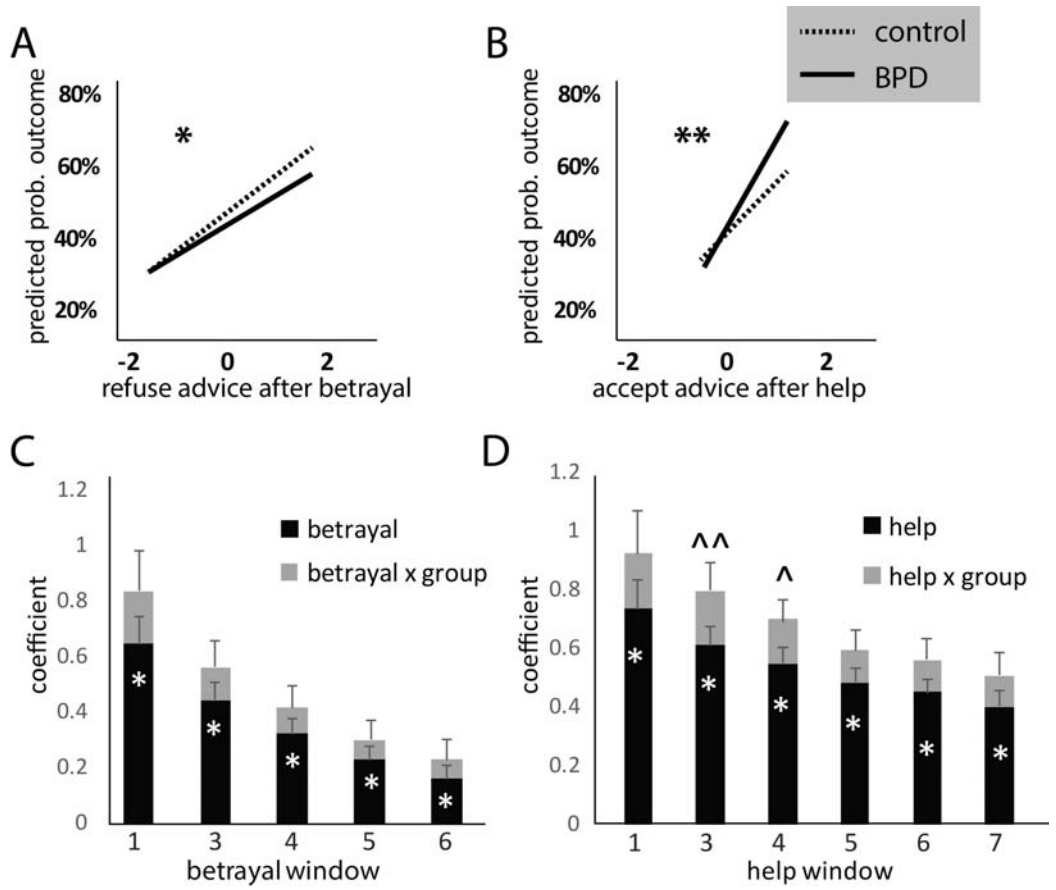

Figure 4: Both betrayal and help impact differently on BPD and control subjects
Predicted probability of outcome differed in BPD versus control subjects exposed to a recent instance of betrayal (incorrect advice) or help (correct advice). BPD subjects were more likely to refuse the advice after betrayal (**A,** * indicates significant effect of group at $p < 0.05$) and take the advice after help (**B,** ** indicates significant group * predictor effect at $p < 0.05$). **C** and **D** depict the results of a closer look at the use of betrayal and help over time. The bar graphs show regression results for a series of time windows, each with at least one betrayal (**C**) or help (**D**) event in the marked number of trials. For example, help window 1 means the advice was correct on the previous trial, help window 4 means the advice was correct on at least one of the previous 4 trials. * predictor $p < 0.05$, ^^ predictor x group $p < 0.05$, ^ predictor x group $p < 0.1$.

All subjects showed diminishing use of these social cues with enlarging time windows; the significant group x predictor interaction in help windows 3 and 4 but not in betrayal windows suggests a slower decrement of help use for decision making in the BPD group compared to the control group. This is not an effect that we observed for the use of betrayal.

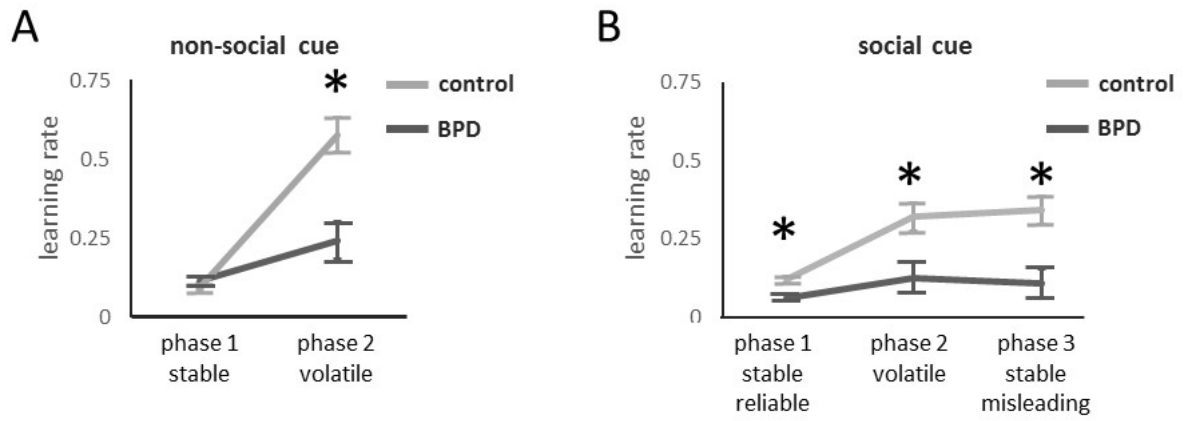

Figure 5. Learning rates were estimated for each individual differences analysed by group and condition. * indicates significant between-group difference. **A)** For learning from the non-social cue (color), there was a group x condition interaction ($F = 19.78$, $p < 0.001$, error df = 1, hypothesis df = 41). Learning was slow and not significantly different by group when reward probability was stable ($t = -1.38$, $p > 0.05$). However, when reward probability became volatile, the BPD group had a blunted response (less of an increase in learning) compared to controls ($t = 4.12$, $p < 0.001$). **B)** For learning from the social cue, a group x condition interaction was also observed ($F = 5.81$, $p < 0.01$, hypothesis df = 2, error df = 40). Here, learning was slower in the BPD group than in the control under all three conditions: stable reliable $t = 4.02$, $p < 0.001$, volatile $t = 2.90$, $p < 0.01$, stable misleading $t = 3.44$, $p < 0.005$), and there was again a blunted response to volatility of reward probability under the volatile condition.

Supplement

Methods and Materials

Self-report scales:

The Borderline Symptom List (BSL-23) is a 23-item scale with established reliability (Cronbach's alpha 0.94-0.96) and validity (r = 0.96 versus the longer BSL-95, r = 0.87 versus the Beck Depression Inventory, and r = 0.48 versus the general psychopathology scale SCL-90) in initial psychometric studies (27). In our sample, 19 control subjects and 20 BPD subjects completed the scale, and Cronbach's alpha is 0.95.

The Beck Anxiety Inventory (BAI) is a 21-item scale is established reliability (Cronbach's alpha 0.94) and validity (r = 0.54 versus diary reports of anxiety) in an initial psychometric validation study (28, 29). In our sample, 20 control subjects and 20 BPD subjects completed the scale, and Cronbach's alpha is 0.95.

The Beck Depression Inventory (BDI-II) is a 21-item scale with established reliability (Cronbach's alpha 0.9) and validity (r = 0.71 – 0.86 versus a range of commonly used depression scales) in a large meta-analysis (30, 31). In our sample, 20 control subjects and 20 BPD subjects completed the scale, and Cronbach's alpha is 0.97.

The Structured Clinical Interview for DSM-IV Personality Questionnaire, though not initially designed as a stand-alone instrument, has been found to have few false-negatives, and was found to be reliable compared to the clinician-administered version (32) (33) (34). In our sample, 20 subjects in each group completed this measure.

Learning model

The authors summarize the model as follows: "The 2-armed bandit model in hBayesDM updates expected value according to:

$$V(t+1) = V(t) + A * (R(t) - V(t))$$

Where V is the expected value on trial t, A is the learning rate, and R is the outcome of the chosen option on trial t. V is only updated for the chosen option. The values (V) are entered into a softmax function with an inverse temperature (Tau) that measures how sensitive subjects are to the differences in expected value between choice options. Higher and

lower values for Tau indicate responding more deterministically and randomly with respect to the learned expected values, respectively." (personal communication with Nathaniel Haines)

The actual implementation is available at

https://github.com/CCS-Lab/hBayesDM/blob/master/exec/bandit2arm_delta.stan

References:

1. Peters E, Lataster T, Greenwood K, Kuipers E, Scott J, Williams S, et al. (2012): Appraisals, psychotic symptoms and affect in daily life. *Psychological medicine*. 42:1013-1023.
2. McIntyre JC, Wickham S, Barr B, Bentall RP (2017): Social Identity and Psychosis: Associations and Psychological Mechanisms. *Schizophrenia bulletin*.
3. Ng R, Fillet P, DeWitt M, Heyman GD, Bellugi U (2015): Reasoning About Trust Among Individuals With Williams Syndrome. *Am J Intellect Dev Disabil*. 120:527-541.
4. Kennedy DP, Glascher J, Tyszka JM, Adolphs R (2009): Personal space regulation by the human amygdala. *Nature neuroscience*. 12:1226-1227.
5. King-Casas B, Sharp C, Lomax-Bream L, Lohrenz T, Fonagy P, Montague PR (2008): The rupture and repair of cooperation in borderline personality disorder. *Science*. 321:806-810.
6. Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008): Associative learning of social value. *Nature*. 456:245-249.
7. Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007): Learning the value of information in an uncertain world. *Nature neuroscience*. 10:1214-1221.
8. Sevgi M, Diaconescu AO, Tittgemeyer M, Schilbach L (2015): Social Bayes: Using Bayesian Modeling to Study Autistic Trait-Related Differences in Social Cognition. *Biological psychiatry*.
9. Brazil IA, Hunt LT, Bulten BH, Kessels RP, de Bruijn ER, Mars RB (2013): Psychopathy-related traits and the use of reward and social information: a computational approach. *Frontiers in psychology*. 4:952.
10. Zanarini MC, Frankenburg F.R., Sickel, A.E., Yong, L. (1996): The Diagnostic Interview for DSM-IV Personality Disorders. . In: Mclean Hospital LftSoAD, editor. Belmont, Mass.
11. Zanarini MC, Gunderson, J.G., Frankenburg F.R., Chauncey, D.L. (1989): The Revised Diagnostic Interview for Borderlines: Discriminating BPD from other axis II disorders. *Journal of personality disorders*. 3:10-18.
12. First MB, Spitzer, R.L., Gibbon, M., Williams, J.B.W. (2007): Structured Clinical Interview for DSM-IV-TR Axis I Disorders-Patient Edition. In: Institute BRDNYSP, editor. New York, New York.
13. Wilkinson GS, Robertson, GJ. (2006): Wide Range Achievement Test 4. Lutz, FL: Psychological Assessment Resources.
14. Uttl B (2002): North American Adult Reading Test: age norms, reliability, and validity. *J Clin Exp Neuropsychol*. 24:1123-1137.
15. Fineberg SK, Leavitt J, Deutsch-Link S, Dealy S, Landry CD, Pirruccio K, et al. (2016): Self-reference in psychosis and depression: a language marker of illness. *Psychological medicine*. 46:2605-2615.
16. Pennebaker JW, Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. Austin, TX (www.liwc.net) (2007): *The development and psychometric properties of LIWC2007. [Software manual].* . Austin, TX.
17. Ahn W-Y, Haines N, Zhang L (2017): Revealing neuro-computational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*. 1:24-57.
18. Fineberg SK, Deutsch-Link S, Ichinose M, McGuinness T, Bessette AJ, Chung CK, et al. (2014): Word use in first-person accounts of schizophrenia. *The British journal of psychiatry : the journal of mental science*.
19. Pennebaker JW (2011): *The secret life of pronouns : what our words say about us*. 1st U.S. ed. New York: Bloomsbury Press.
20. Bertsch K, Gamer M, Schmidt B, Schmidinger I, Walther S, Kastel T, et al. (2013): Oxytocin and reduction of social threat hypersensitivity in women with borderline personality disorder. *The American journal of psychiatry*. 170:1169-1177.
21. Schulze L, Schmahl C, Niedtfeld I (2016): Neural correlates of disturbed emotion processing in borderline personality disorder: a multimodal meta-analysis. *Biological psychiatry*. 79:97-106.
22. Bartz J, Simeon D, Hamilton H, Kim S, Crystal S, Braun A, et al. (2011): Oxytocin can hinder trust and cooperation in borderline personality disorder. *Social cognitive and affective neuroscience*. 6:556-563.
23. Paret C, Hoesterey S, Kleindienst N, Schmahl C (2016): Associations of emotional arousal, dissociation and symptom severity with operant conditioning in borderline personality disorder. *Psychiatry research*. 244:194-201.
24. Berlin HA, Rolls ET, Iversen SD (2005): Borderline personality disorder, impulsivity, and the orbitofrontal cortex. *The American journal of psychiatry*. 162:2360-2373.

25. Stevens A, Burkhardt M, Hautzinger M, Schwarz J, Unckel C (2004): Borderline personality disorder: impaired visual perception and working memory. *Psychiatry research*. 125:257-267.
26. Hula A, Vilares I, Lohrenz T, Dayan P, Montague PR (2018): A model of risk and mental state shifts during social interaction. *PLoS computational biology*.
27. Bohus M, Kleindienst N, Limberger MF, Stieglitz RD, Domsalla M, Chapman AL, et al. (2009): The short version of the Borderline Symptom List (BSL-23): development and initial data on psychometric properties. *Psychopathology*. 42:32-39.
28. Fydrich T, Dowdall D, Chambless DL (1992): Reliability and Validity of the Beck Anxiety Inventory. *Journal of Anxiety Disorders*. 6:55-61.
29. Beck AT, Epstein N, Brown G, Steer RA (1988): An inventory for measuring clinical anxiety: psychometric properties. *Journal of consulting and clinical psychology*. 56:893-897.
30. Wang YP, Gorenstein C (2013): Psychometric properties of the Beck Depression Inventory-II: a comprehensive review. *Revista brasileira de psiquiatria*. 35:416-431.
31. Steer RA, Ball R, Ranieri WF, Beck AT (1999): Dimensions of the Beck Depression Inventory-II in clinically depressed outpatients. *J Clin Psychol*. 55:117-128.
32. Ball SA, Rounsaville BJ, Tennen H, Kranzler HR (2001): Reliability of personality disorder symptoms and personality traits in substance-dependent inpatients. *Journal of abnormal psychology*. 110:341-352.
33. Ekselius L, Lindstrom E, von Knorring L, Bodlund O, Kullgren G (1994): SCID II interviews and the SCID Screen questionnaire as diagnostic tools for personality disorders in DSM-III-R. *Acta psychiatrica Scandinavica*. 90:120-123.
34. Jacobsberg L, Perry, S., Frances, A. (1995): Diagnostic agreement between the SCID-II screening questionnaire and the Personality Disorder Examination. *J Pers Assess*. 65:428-433.