# Manuscript Details

| | |
|---|---|
| **Manuscript number** | JPSYCHIATRRES_2017_279 |
| **Title** | Aberrant link between empathy and social attribution style in borderline personality disorder |
| **Article type** | Original article |

### Abstract

In social interactions, we often need to quickly infer why other people do what they do. More often than not, we infer that behavior is a result of personality rather than circumstances. It is unclear how the tendency itself may contribute to psychopathology and interpersonal dysfunction. Borderline personality disorder (BPD) is characterized by severe interpersonal dysfunction. Here, we investigated if this dysfunction is related to the tendency to over-attribute behaviors to personality traits. Healthy controls and patients with BPD judged positive and negative behaviors presented within a situational constraint, during functional magnetic resonance imaging. Before the experiment, we measured trait levels of empathy, paranoia, and need for cognition. Behaviorally, we found that empathy levels predicted the tendency to attribute behavior to traits in healthy controls, whereas in patients with BPD this relationship was significantly weakened. Whole brain analysis of group-by-empathy interaction revealed that when participants judged the behavior during the attribution phase, several brain regions implicated in mentalizing distinguished patients from controls: In healthy controls, neural activity scaled negatively with empathy, but this relationship was reversed in BPD patients. Due to the cross-sectional study design we cannot establish a causal link between empathy and social attributions. These findings indicate that self-reported tendency to feel for others is related to the tendency to integrate situational information beyond personality. In BPD patients, by contrast, the association between empathy and attribution was significantly weaker, rendering empathy less informative in predicting the overall attribution style.

| | |
|---|---|
| **Keywords** | Empathy; Borderline personality disorder; social cognition; attribution; temporoparietal junction; mentalizing |
| **Taxonomy** | Causal Judgment, Functional Magnetic Resonance Imaging |
| **Manuscript region of origin** | North America |
| **Corresponding Author** | Daniela Schiller |
| **Order of Authors** | Philipp Homan, Marianne Reddan, Tobias Brosch, Harold Koenigsberg, Daniela Schiller |
| **Suggested reviewers** | Stefan Roepke, Barbara Stanley, Sabine Herpertz, Brian Haas |

## Submission Files Included in this PDF

**File Name [File Type]**

fae-bpd_coverletter.doc [Cover Letter]

fae-bpd_responses.docx [Response to Reviewers]

fae-bpd_ms_abstract.doc [Abstract]

fae-bpd_ms_titlepage.doc [Title Page (with Author Details)]

fae-bpd_ms.docx [Manuscript File]

fae-bpd_ms_coi.doc [Conflict of Interest]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Mount Sinai

***Daniela Schiller, PhD***
Associate professor
Department of Psychiatry
Department of Neuroscience

One Gustave L. Levy Place, Box 1230
New York, NY 10029
Tel. (212) 824-8977
Fax. (212) 731-7804
Email: daniela.schiller@mssm.edu

May 25, 2017

Dear Editor,

Please find below the revised manuscript titled "*Aberrant link between empathy and social attribution style in borderline personality disorder*", which we submit as Original Article for publication in *the Journal of Psychiatric Research.*

Thank you for considering our manuscript and giving us the opportunity to reply to the reviewers' comments. Below we provide point-by-point responses, addressing the reviewers' concerns in full and improving the clarity of the manuscript.

Given the novelty of this report as a first examination of empathy and social attribution style in borderline personality disorder, we hope that you agree that the manuscript is worth pursuing publication in *the Journal of Psychiatric Research*.

Sincerely,
Philipp Homan, MD, PhD and Daniela Schiller, PhD

# Response to Reviews

## Review #1

**Comment**

This paper reports on a small FMRI study of the relationship between attribution style (personality trait or situation), empathy measured by self-report, and brain mechanisms associated with mentalization across healthy comparison subjects and patients with borderline personality disorder. It was found that the groups did not differ in their overall tendency to attribute behavior to traits. With increasing empathy, healthy controls were more likely to overcome the tendency to attribute behavior to personality traits. In borderline patients, this relationship was significantly weakened. In healthy controls, neural activity scaled negatively with empathy but this relationship was reversed in borderline patients.

The concepts being studied are interesting but the paper does not easily and cleanly take the reader from one section to another. In addition, the paper has a large number of comparisons and it not clear that they were corrected for multiple comparisons. In terms of other issues, the results of three self-report measures were reported in Table 1, although these measures were not mentioned in the Methods section of the paper.

- **Reply:** We thank the reviewer for this comment. In addition to the self report measures mentioned on page 5, we added the Interpersonal Reactivity Index: "Before the experiment, measures of paranoia (Freeman, 2005), empathy (Mayer et al., 1999), interpersonal reactivity (Davis 1980) and the need for cognition (Cacioppo and Petty, 1982) were taken."

  Regarding multiple comparisons, correcting for the 3 models tested on trait attributions with a Bonferroni-correction, the resulting *P*-value threshold was $P < 0.017$ and the interaction of group-by-empathy on trait attributions with $P = 0.04$ was not significant anymore. However, as we were testing for more subtle influences here with interactions, the risk of missing a potentially real effect (type II error) and the risk of falsely assuming an effect that was driven by chance (type I error) should be equally considered. In addition, our neuroimaging findings do reflect the interaction we found on the behavioral level and do survive a whole-brain correction for family-wise error (consistently on the cluster-level, in part even on the voxel level), suggesting an acceptable rate of type I error. We now report results before and after Bonferroni-correction in the paper.
- **Action taken:** We added the corresponding information and have also revised the language throughout to improve the clarity of the manuscript.

**Comment**

In the Introduction, it is stated that BPD is an enduring condition with a suicide rate of 10%. Recent findings suggest that neither of these statement is accurate.

- **Action taken:** Following the reviewer's suggestion, we clarified that the suicide rate in BPD is around 8% according to a more recent meta-analysis (Pompili et al., 2005, Nord J Psychiatry) and removed the expression "enduring".

**Comment**
Finally, the number of figures could be reduced and the content laid out in a more easily understood style.

- **Action taken:** Since removing some visualization of the results might reduce the clarity of the manuscript we propose to keep the current figures but we have modified some language throughout to improve clarity.

## Review #2

**Comment**
In this interesting and timely report, Homan and colleagues show that self-reported empathy levels predict the tendency to attribute behavior to traits rather than circumstances, and that this relationship is significantly weaker in patients with borderline personality disorder. In healthy controls, BOLD signal activity in a “mentalizing network” during the attribution phase of the task was inversely correlated with empathy, whereas in borderline personality disorder patients, this relationship was reversed. There is much to like about this study. It is a timely addition to a relatively limited literature on the neurobiological substrates of mentalizing, especially in borderline personality disorder. The task is clever and well designed. And the manuscript is clear and well written. I have just a few suggestions for minor revisions:

Self-reported empathy predicted tendency to attribute behavior to traits in healthy controls and this relationship was weakened in controls. To what extent does this have to do with the relationship between empathy and attribution vs. differences in patients’ ability to accurately assess their own levels of empathy? The authors discuss this briefly in the last paragraph of the discussion, pointing out that this explanation is unlikely because BPD patients accurately self-report higher levels of paranoia. However, a tendency to over-estimate self-reported positive traits could be distinct from accurately self-reporting negative traits. Can the authors point to any evidence from the existing literature that can speak to this question?

- **Reply:** We thank the reviewer for this insightful comment. We are aware of at least one study that may indicate the possibility of enhanced self-insight in patients with BPD (Flury et al., 2008, Journal of Research in Personality, 42, 312–332). That study reported significantly increased empathic accuracy in individuals with high-borderline compared to low-borderline features. However, another study found that patients with BPD "dramatically overestimated their levels of Agreeableness and Conscientiousness, estimating themselves to be slightly above average on each of these characteristics but actually scoring well below average on both" (Morey 2014, Borderline Personality Disorder and Emotion Dysregulation, 1:4). Given this evidence, we decided to remove the potential sentence in the Discussion as a introspection deficit can indeed not be ruled out for positive traits.

- **Action taken:** We removed the following from the Discussion: ~~However, the results of the baseline ratings speak against this explanation as levels of self-reported paranoia were higher compared to controls, which was expected and in line with previous literature suggesting that BPD patients tend to perceive others as negative and malevolent (reviewed in Roepke et al., 2013). We have thus no reason to believe in distortions in the response patterns.~~

**Comment**
One of the main strengths of this work is that it adds to a very limited neuroimaging literature in borderline personality disorder. Some additional (brief) discussion of what is known from this literature and how it relates to the authors' findings could be useful for readers.

- **Action taken:** We included a paragraph on neuroimaging findings in BPD in the Discussion (p. 15f.): "Previous neuroimaging studies in patients with BPD have focused primarily on emotion processing and emotion regulation. Main findings include increased neural response of limbic regions and diminished recruitment of frontal brain regions implicated in emotional regulation (Krause-Utz et al., 2014). A recent meta-analysis found that patients consistently showed increased activation of the left amygdala and posterior cingulate cortex and attenuated activity of the bilateral dorsolateral prefrontal cortex during the processing of negative emotional stimuli (Schulze et al., 2016). The current study extends these findings by focusing on social-cognition and by taking empathy into account."

**Comment**
Given concerns about motion-related artifact in the BOLD signal, it would be helpful to include some descriptive statistics about head motion in the healthy controls vs. BPD patients, especially since there is very limited published data on this issue in BPD.

- **Reply:** Following the reviewer's suggestion, we calculated the total head movement in mm per participant during the fMRI scan, following the algorithm described in Savalia et al. (2017, Human Brain Mapping 38:472-492). This algorithm calculates the sum of the absolute values of the six differentiated realignment parameters at each frame of an individual fMRI scan; this sum thus reflects the total movement as a positive displacement value. Healthy controls and patients with BPD did not differ significantly at total head movement (HC: mean = 220.24 mm, SD = 217.37 mm; BPD: mean = 166.81 mm, SD = 89.39 mm; $t(21.3) = 0.94$, $P = 0.36$).
- **Action taken:** We included this parameter in the Neuroimaging Results section (p. 14).

**Comment**
4. There is a typo in the legend for Figure 1, line 4.

- **Action taken:** We corrected the typo.

## Review #3

I thought this paper is fine as is.  It was well done, has interesting and relevant findings, and I do not have any suggestions for the authors.

- **Reply:** We thank the reviewer for this positive input.

# Abstract

In social interactions, we often need to quickly infer why other people do what they do. More often than not, we infer that behavior is a result of personality rather than circumstances. It is unclear how the tendency itself may contribute to psychopathology and interpersonal dysfunction. Borderline personality disorder (BPD) is characterized by severe interpersonal dysfunction. Here, we investigated if this dysfunction is related to the tendency to over-attribute behaviors to personality traits. Healthy controls and patients with BPD judged positive and negative behaviors presented within a situational constraint, during functional magnetic resonance imaging. Before the experiment, we measured trait levels of empathy, paranoia, and need for cognition. Behaviorally, we found that empathy levels predicted the tendency to attribute behavior to traits in healthy controls, whereas in patients with BPD this relationship was significantly weakened. Whole brain analysis of group-by-empathy interaction revealed that when participants judged the behavior during the attribution phase, several brain regions implicated in mentalizing distinguished patients from controls: In healthy controls, neural activity scaled negatively with empathy, but this relationship was reversed in BPD patients. Due to the cross-sectional study design we cannot establish a causal link between empathy and social attributions. These findings indicate that self-reported tendency to feel for others is related to the tendency to integrate situational information beyond personality. In BPD patients, by contrast, the association between empathy and attribution was significantly weaker, rendering empathy less informative in predicting the overall attribution style.

# Aberrant link between empathy and social attribution style in borderline personality disorder

Philipp Homan[1], Marianne C. Reddan[2], Tobias Brosch[3], Harold W. Koenigsberg[1,4], Daniela Schiller[1,5*]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2]Department of Psychology, University of Boulder, Colorado, USA

[3]Department of Psychology, University of Geneva, Switzerland

[4]James J Peters VA Medical Center, Bronx, NY, USA

[5]Department of Neuroscience, and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

*To whom correspondence should be addressed; E-mail: daniela.schiller@mssm.edu

Abstract: 246 words
Body: 3599 words
Figures: 5
Tables: 3
Supplemental information: 0

**Keywords:** Empathy; Borderline personality disorder; social cognition; attribution; temporoparietal junction; mentalizing

# Aberrant link between empathy and social attribution style in borderline personality disorder

Philipp Homan[1], Marianne C. Reddan[2], Tobias Brosch[3], Harold W. Koenigsberg[1,4], Daniela Schiller[1,5*]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[2]Department of Psychology, University of Boulder, Colorado, USA

[3]Department of Psychology, University of Geneva, Switzerland

[4]James J Peters VA Medical Center, Bronx, NY, USA

[5]Department of Neuroscience, and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

*To whom correspondence should be addressed; E-mail: daniela.schiller@mssm.edu

Abstract: 246 words
Body: 3599 words
Figures: 5
Tables: 3
Supplemental information: 0

**Keywords:** Empathy; Borderline personality disorder; social cognition; attribution; temporoparietal junction; mentalizing

# Abstract

In social interactions, we often need to quickly infer why other people do what they do. More often than not, we infer that behavior is a result of personality rather than circumstances. It is unclear how the tendency itself may contribute to psychopathology and interpersonal dysfunction. Borderline personality disorder (BPD) is characterized by severe interpersonal dysfunction. Here, we investigated if this dysfunction is related to the tendency to over-attribute behaviors to personality traits. Healthy controls and patients with BPD judged positive and negative behaviors presented within a situational constraint, during functional magnetic resonance imaging. Before the experiment, we measured trait levels of empathy, paranoia, and need for cognition. Behaviorally, we found that empathy levels predicted the tendency to attribute behavior to traits in healthy controls, whereas in patients with BPD this relationship was significantly weakened. Whole brain analysis of group-by-empathy interaction revealed that when participants judged the behavior during the attribution phase, several brain regions implicated in mentalizing distinguished patients from controls: In healthy controls, neural activity scaled negatively with empathy, but this relationship was reversed in BPD patients. Due to the cross-sectional study design we cannot establish a causal link between empathy and social attributions. These findings indicate that self-reported tendency to feel for others is related to the tendency to integrate situational information beyond personality. In BPD patients, by contrast, the association between empathy and attribution was significantly weaker, rendering empathy less informative in predicting the overall attribution style.

# 1. Introduction

Functional and healthy relationships rely on specific social skills. In social interactions, we often need to quickly infer why other people do what they do. Sometimes we infer that behavior is a result of personality rather than circumstances. In other words, we attribute behavior to traits rather than the context, a cognitive bias known as the Fundamental Attribution Error (Ross, 1977) or Correspondence Bias (Gilbert and Malone, 1995). Overcoming this tendency appears to require motivation and cognitive effort (Trope, 1986), but it is unclear how the tendency itself may contribute to psychopathology and interpersonal dysfunction.

A significant personality disorder that is characterized by interpersonal dysfunction is borderline personality disorder (BPD). Patients with BPD suffer from rapid alternations of intense attachment and intense detachment in interpersonal situations, extreme devaluation or idealization of others, and impairments in mentalizing, the capacity to understand other people's behavior in terms of their likely feelings, desires and goals (Gunderson, 1996; Gunderson and Lyons-Ruth, 2008; Koenigsberg et al., 2009; Fonagy and Bateman, 2008; American Psychiatric Association, 2013). BPD is a severe condition, present in an estimated 2.7% of the population (Tomko et al., 2013), and characterized by emotional instability and impulsive aggression (Skodol et al., 2002). With a suicide rate of around 8% (Pompili et al., 2005), unfavorable outcomes are common in BPD and appear to be closely related to interpersonal dysfunction (Koenigsberg et al., 2001).

Recent efforts have tried to uncover the neural correlates underlying this dysfunction. Key findings include reduced activity in the left superior temporal cortex when inferring others' mental states together with increased responsiveness of the right mid-insula and left posterior insula when sharing others' emotions (Dziobek et al., 2011; Mier et al., 2013; Roepke et al., 2013). In addition, activity in the temporoparietal junction and superior temporal sulcus scaled negatively with borderline symptoms in an emotional perspective-taking task (Haas and Miller, 2015). Together, these findings suggest hypofunction in core mentalizing regions in BPD during social cognition tasks.

Here, we asked whether the tendency to attribute behavior to personality traits rather than situational context differs in BPD patients compared to healthy neurotypical control participants; whether relevant personality variables, including empathy, suspiciousness, and need for cognition, differentially mediate social attribution styles in this population; and what are the neural mechanisms underlying apparent differences. To address these questions, healthy controls and BPD patients performed a social attribution and evaluation task during functional neuroimaging. The task involved judging people's positive and negative behaviors within a situational constraint, which allowed us to test the tendency to attribute behavior to traits as a function of behavior valence (Fig. 1).

# 2. Material and methods

## 2.1. Study design

We recruited healthy human participants between 18 and 65 years of age, ethnicity, education-level and sex-matched patients with borderline personality disorder. Participants in the borderline group met DSM-IV criteria for borderline personality disorder and were currently under treatment. Healthy participants did not meet DSM-IV criteria for any axis I or axis II disorder. Diagnostic assessments were obtained using the Structured Clinical Interview for DSM-IV–Patient Edition and the Structured Clinical Interview for DSM-IV Axis II Personality Disorders. Our group achieved an inter-rater reliability of 0.81 for diagnosing borderline personality disorder.

All participants provided written informed consent and were financially compensated for their participation. The Institutional Review Board of the Icahn School of Medicine at Mount Sinai approved the experiment. This was a between-subjects experimental design on a single day. Participants performed a social judgment task during a scan of functional magnetic resonance imaging. Before the experiment, measures of paranoia (Freeman, 2005), empathy (Mayer et al., 1999), interpersonal reactivity (Davis 1980) and the need for cognition (Cacioppo and Petty, 1982) were taken. The empathy scale measured the affective dimension of empathy including empathic suffering, sharing of positive emotions, and emotional attention. With the paranoia scale we assessed suspiciousness, i.e., assumptions such as that friends, acquaintances or strangers might be hostile. The need for cognition scale measured the tendency to take and enjoy cognitive effort when it is required.

## 2.2. Social judgment paradigm

### 2.2.1. Stimuli

For the social judgment task, we used 32 vignettes that consisted of scenarios describing the behavior of named characters in a given situation (Brosch et al., 2013). The behavioral segment described a certain behavior (e.g., "Mike left the restaurant in a hurry without tipping the waitress"), whereas the situational segment described the circumstances under which the behavior took place (e.g., "Mike's baby was screaming"). One half of the vignettes described a positive behavior, the other half a negative behavior. Notably, each situational segment was constructed in a way that could potentially relativize both the positive and the negative behaviors. Behavioral and situational segments were presented separately, and the presentation order was counterbalanced across scenarios (there was no effect of order; $P = 0.5$).

### 2.2.2. Behavioral task

During two consecutive runs, participants were asked to read brief vignettes about male or female persons, consisting of a behavior segment, describing a positive or negative behavior, and a situation segment, describing the circumstances surrounding the behavior[19] (Fig. 1). Each segment was shown for 6 seconds and separated by a fixation cross that was shown for an intertrial interval varying between 2 and 6 seconds. Participants were then presented with 2 consecutive rating screens for a maximum of 10.5 seconds where they were instructed to rate (see next section) whether the behavior was due to the person's personality (i.e., "dispositional attribution") or the circumstances (i.e., "situational attribution"), as well as how much they liked the person. Each rating screen was replaced by a feedback screen ("Thank You") for 500 ms after the judgment was made, and a final feedback screen ("Thank you. Your responses were recorded") after the last rating screen appeared for the remainder of a maximum 10.5 seconds period. The character’s face (with neutral expression) was presented continuously throughout the information, rating, and feedback screens. Finally, a fixation cross was shown for an intertrial interval varying between 2 and 6 seconds.

The experimental order was pseudo-randomized and counterbalanced across participants so that the order of screens differed between participants; that is, across all trials, there were vignettes that started with a behavioral segment and others with a situational segment; the order of the rating screens also differed across trials as well as the assignment of vignettes to the different faces. E-Prime 2.0 (Psychology Software Tools Inc., Pittsburgh, PA) was used as presentation software.

### 2.2.3. Behavioral assessment

During the task, participants rated the causation of the behavior and liking of the person trial by trial on a discrete visual analogue scale between 1 and 8 (1 indicating "not at all", 8 indicating "extremely" for the liking ratings, while 1 indicated "situational factors" and 8 indicated "dispositional factors" for the attribution).

### 2.2.4. fMRI data acquisition

A 3 T Philips Gemini scanner and Philips standard head coil were used for data acquisition. Functional images were recorded in two consecutive scanning sessions. The sessions comprised 642 volumes each. A single-shot gradient echo EPI sequences (TR = 2.0 s; TE = 25 ms; FoV = 192 cm, flip angle = 75°) was used to obtain 46 oblique-axial slices with 2 mm thickness, a 1 mm inter-slice gap, and an in-plane resolution of 3 x 3 mm parallel to the anterior commissure-posterior commissure line.

### 2.2.5. Behavioral data analysis

The primary outcome measure was the number of trait attributions when exposed to positive and negative behaviors. Attributions score of 5 – 8 were considered dispositional attributions whereas scores of 1 – 4 were considered situational attributions. We calculated the fraction of trials where a dispositional attribution was recorded out of the total number of trials. For each subject, this fraction was calculated separately for positive and negative behavior trials. Our analysis was framed as a test for a significant main effect of group for potential differences in overall tendency to perform trait attributions, and interaction of group-by-valence, potentially reflecting more trait attributions for negative or positive behaviors by one of the groups.

The liking ratings (mean scores on the 1-8 scale) were assessed for overall group differences, as well as their correspondence to trait attributions. That is, whether liking scores were higher following trait attribution for positive behavior and lower following trait attribution for negative behavior compared to liking when situational attributions had been made. Lastly, analyses of covariance, including effects of paranoia, empathy and need for cognition, as well as their interaction with group were calculated separately for the outcome measures attribution and liking. Thus, we tested for group differences in the linear relationship between each covariate and outcome measure (Miller et al., 2001). Statistically, this corresponds to testing whether the difference in slopes of each covariate was significantly different from zero. The statistical significance threshold was set at alpha = 0.05, two-tailed. The statistical software package R (R Core Team, 2016) was used for all analyses.

### 2.2.6. fMRI data analysis

Functional data were analyzed with the SPM12 software package (Wellcome Trust Centre for Neuroimaging, London, UK) and the toolboxes marsbar. Native-space images were first realigned, slice-time corrected, and coregistered to each subject's structural scan. Structural image preprocessing included segmentation, bias correction, and spatial normalization; these normalization parameters were also used to normalize the functional images. Finally, functional images were smoothed with a Gaussian kernel (8 mm FWHM) and resampled to 2 x 2 x 2 mm voxels.

Data were modeled voxelwise with a general linear model (GLM) for each of the participants. Both runs were included in one GLM as separate sessions and thus separate regressors. GLM regressors accounted for behavioral and situational segments as well as subjects' attribution and liking ratings for each stimulus separately. Durations for attribution and liking ratings were modeled individually using the participants' reaction times, as this method provides the most sensitivity for decision-making tasks (Grindband et al., 2008). Trials where no response was recorded were modeled in an additional error-regressor. Each regressor was convolved with the canonical hemodynamic response function provided by SPM12, and a high-pass filter with a cutoff period of 128 s and an autoregressive first order model correction for temporal autocorrelation was applied. Additional regressors were included as parametric modulators for each of the aforementioned regressors to account for the valence difference of the behavior vignettes (positive or negative). Finally, six regressors modeling affine head-motion parameters were included as additional covariates in all GLMs. After computing contrast images for each subject, group analyses assessed random-effects across all subjects by calculating a second-level analysis of covariance and testing for both positive and negative blood oxygen level dependent (BOLD) effects.

The contrast of interest was the effect of attribution, and we were interested how the predictors empathy, paranoia or need for cognition differed between groups in its relationship with the outcome measure, i.e., the neural activity during attribution. Thus, statistically we tested for differences of the slopes of empathy, paranoia or need for cognition between groups, i.e., the group-by-covariate interactions across the whole brain. We report any responses that survived a whole-brain correction for family-wise error at the cluster level (*P* < 0.05), with a threshold at the voxel level of *P* < 0.001 to ensure that the random field assumptions were met and cluster level inferences were valid (Friston, 2009; Eklund et al., 2016; Flandin and Friston, 2016).

# 3. Results

## 3.1. Sample characteristics

There were 17 healthy controls cases (HCC) participants and 18 BPD patients enrolled in the study. Fifteen controls and 14 patients completed the full experiment and the remaining subjects could thus not be included in the final analysis due to missing data. Baseline characteristics are given in Table 1. The groups did not differ in age or sex; significant increases were evident in BPD patients in measures of psychopathology including depression, stress and paranoia. These measures were overall in line with the expected symptomatology (Tomko et al., 2013). The groups did not differ in need for cognition but BPD patients showed lower levels of empathy (Table 1).

## 3.2. Behavioral results

The primary behavioral outcome measure was the number of trait attributions when exposed to positive and negative behaviors. Overall, the BPD patients versus healthy controls did not significantly differ when attributing negative or positive behavior to traits (no group-by-valence interaction on trait attributions; *F* (1, 32) = 2.34, *P* = 0.135). Both groups made more trait attributions when judging positive compared to negative behavior (valence effect; *F* (1, 32) = 19.48, *P* = 0.0001) but they did not differ in their overall tendency to attribute behavior to traits (group effect; *F* (1, 56.68) = 0.27, *P* = 0.61; Fig. 2).

The groups also did not differ in liking ratings (no group-by-valence interaction; *F* (1, 33.189) = 1.64, *P* = 0.21; no group effect; *F* (1, 33.095) = 1.23, *P* = 0.28). Confirming task validity, the participants rated positive behaviors overall higher than negative behaviors (valence effect; *F* (1, 33.189) = 13.35, *P* = 0.0009), and the liking ratings in both groups corresponded to the valence of the trait attributions (valence-by-attribution interaction; *F* (1, 27.89) = 15.92, *P* = 0.0004; Fig. 3) when categorizing the trials by the type of attribution that had been made. Specifically, when a dispositional judgment had been made, liking scores for positive behaviors were significantly higher then for negative behaviors (*t* (36.43) = 5.69, *P* < 0.0001), but no significant difference was observed when the situational context was considered (*t* (27.89) = -0.87, *P* = 0.39). Furthermore, when comparing between attribution types within valence, liking scores increased for negative behavior (*t* (32.01) = 2.19, *P* = 0.036) and decreased for positive behavior (*t* (28.9) = 3.26, *P* = 0.003) when the situational context was considered compared to when a dispositional judgment had been made.

Finally, there were no differences in response times evident for trait attributions (no group-by-valence interaction; *F* (1, 220.071) = 0.186, *P* = 0.67; no group effect; *F* (1, 33.06) = 1.08, *P* = 0.31) or liking ratings (no group-by-valence interaction; *F* (1, 132.498) = 0.20, *P* = 0.65; no group effect; *F* (1, 33.256) = 0.69, *P* = 0.41), respectively. Taken together, these results indicate that controls and BPD patients did not differ in trait attributions. Both groups overall made more trait attributions when judging positive behaviors and adjusted their liking ratings according to their trait attributions.

Next, to uncover the possibly more subtle influences of personality variables on social attribution, we focused on how empathy, paranoia and need for cognition moderated the process of attribution and its neural correlates. In HCC, we found that empathy levels predicted the tendency to attribute behavior to traits. With increasing empathy, HCC were more likely to overcome the tendency to attribute behavior to personality traits. In patients with BPD this relationship was significantly weakened (Table 2 and Fig. 4). The corresponding interactions were not significant when testing for paranoia or need for cognition (Table 2 and Fig. 4).

Correcting these tests for the 3 multiple comparisons made with a Bonferroni-correction, the resulting *P*-value threshold was $P < 0.017$ and the interaction of group-by-empathy on trait attributions with $P = 0.04$ was not significant anymore. However, as we were testing for more subtle influences here with interactions, the risk of missing a potentially real effect (type II error) and the risk of falsely assuming an effect that was driven by chance (type I error) should be equally considered. In addition, the neuroimaging results (see below) reflect the interaction we found here on the neural level, and do survive a whole-brain correction for family-wise error (consistently on the cluster-level, in part even on the voxel level), suggesting an acceptable rate of type I error. Similar analyses of covariance on liking ratings yielded no significant results.

## 3.3. Neuroimaging results

When participants judged the behavior during the attribution phase, several brain regions implicated in mentalizing robustly distinguished patients from controls in their relationship with empathy. Specifically, whole brain analysis of group-by-empathy interaction during the attribution phase revealed activity in the left precuneus, right medial frontal gyrus (MFG), dorsal part of the anterior cingulate cortex (dACC) and the right temporoparietal junction (rTPJ) and left superior temporal gyrus (STG) decreased with increasing empathy in controls, while this effect was reversed in patients with BPD: increasing empathy predicted increasing brain activity (Table 3 and Fig. 5). No significant relationships were found for paranoia or need for cognition.

These findings suggest that activity in key regions of the brain's mentalizing network (Van Overwalle and Baetens, 2009; Lavoie et al., 2016) and their relationship with empathy dissociated BPD patients from controls during the actual attribution process, pointing at an opposite neural recruitment of prominent mentalizing regions that is correlated with empathy.

The more empathic were the healthy controls, the less they activated brain regions related to mentalizing when performing social attributions. In contrast, the more empathic were the BPD patients, the more they activated brain regions related to mentalizing when performing social attributions.

Given concerns about motion-related artifact in the BOLD signal, and potential differences between the BPD and healthy control group, we calculated the total head movement in mm per participant during the fMRI scan. Specifically, we used an algorithm that calculates the sum of the absolute values of the six differentiated realignment parameters at each frame of an individual fMRI scan (Savalia et al., 2017); this sum thus reflects the total movement as a positive displacement value. Healthy controls and patients with BPD did not differ significantly at total head movement (HC:  mean = 220.24 mm, SD = 217.37 mm; BPD: mean = 166.81 mm, SD = 89.39 mm; $t(21.3) = 0.94$, $P = 0.36$).

# 4. Discussion

This study found that empathy informs behavioral attributions of neurotypical individuals to a significantly higher degree than of patients with BPD: The tendency to attribute behavior to traits rather than context was predicted by empathy. This suggests that the self-reported tendency to feel for others may also predict the willingness to see the full picture instead of just the personality when their behavior was judged. In BPD patients, by contrast, the association between empathy and attribution was significantly weaker, rendering empathy less informative in predicting the overall attribution style.

When examining the overall tendency for trait attribution, neither healthy controls not BPD patients showed any bias in trait attributions. Previous research has shown that bias in trait attributions can be induced using various manipulations such as social groups affiliations (e.g., in-group/out-groups or levels of similarity), cognitive load (reviewed in Gawronski and Creighton, 2013), and mood (Forgas, 1998). By contrast, our task did not include a bias-inducing manipulation and was designed as to balance trait attributions and situational judgments in healthy humans (Brosch et al., 2013). This allowed us to examine whether BPD psychopathology may bias patients' trait attributions. Our results indicate no such bias but rather a more subtle alteration of the modulatory role of empathy on these attributions. Yet, it is possible that BPD patients would differ from controls in their susceptibility to bias-inducing manipulations in trait attributions, a possibility that remains open for future studies.

Empathy can be defined as the ability to understand and share the feelings of others (Haas and Miller, 2015), and as such lies at the heart of social cognition (Roepke et al., 2013). Sharing the feelings of others may be seen as the spontaneous dimension of empathy, and partly for this reason, it is also what has lead some authors to question the supposedly positive view on empathy

(Rifkin, 2010) altogether (Bloom, 2013; Lamm and Majdanduic, 2015; Bloom, 2017). According to this skeptical view, the inherent in-group bias of empathy implies that we feel empathy toward people only when we perceive them as close and familiar (Bloom, 2013; Bloom, 2017). Social attribution studies tell us that we also judge in-group members differently: we are more willing to see the full picture when judging their behavior, instead of merely attributing behavior to their personalities (Trope and Liberman, 2010; Stephan et al., 2011; Rim et al., 2009; Nussbaum et al., 2003; Bar-Anan et al., 2006). This suggests a close relationship between empathy and social attribution – the extent to which the emotions of others are shared should then be reflected in the way they are perceived – which our results now directly demonstrate.

Previous findings regarding empathy in BPD have been conflicting: some studies showed that patients exhibit less empathy compared to controls (Minzenberg et al., 2006; Preißler et al., 2010; Fertuck et al., 2009), while others found evidence for heightened empathy (Franzen et al., 2011; Dinsdale and Crespi, 2013) in patients. Although this conflicting pattern has been explained by referring to the different domains of empathy that show reduced cognitive empathy in BPD patients with unchanged or even heightened affective empathy (reviewed in Dinsdale and Crespi, 2013), we here found that even the affective domain of empathy may be reduced, possibly reflecting a compensatory response in order to protect from emotional contagion through the emotions of others that has been found in BPD (Dinsdale and Crespi, 2013).

The neuroimaging results showed a reversed relationship between empathy and neural recruitment in prominent mentalizing regions during social attribution, as activity in these regions scaled positively with empathy in BPD and negatively in controls. Previous neuroimaging studies in patients with BPD have focused primarily on emotion processing and emotion regulation. Main findings include increased neural response of limbic regions and diminished recruitment of frontal brain regions implicated in emotional regulation (Krause-Utz et al., 2014). A recent meta-analysis found that patients consistently showed increased activation of the left amygdala and posterior cingulate cortex and attenuated activity of the bilateral dorsolateral prefrontal cortex during the processing of negative emotional stimuli (Schulze et al., 2016).

Recent work on empathy and BPD used perspective-taking task (Derntl et al., 2010) and found that activity in the right TPJ and superior temporal sulcus decreased with self-reported BPD traits (Haas and Miller, 2015). Although measured in healthy controls, that finding adds to previous work that established a link between the TPJ and empathy as well as theory of mind (Saxe and Kanwisher, 2003). Interestingly, disruptions of the TPJ promote the tendency of falsely attributing hostility to other people (Giardina et al., 2011), a tendency also found in paranoid personality traits that may accompany BPD. Structural abnormalities within the TPJ have been found in female BPD patients and include smaller right compared to left parietal lobes (Irle et al., 2005), while functional studies highlight the TPJ's role in the cognitive aspects of empathic processing (Haas and Miller, 2015), suggesting that reduced TPJ volume and function may be a neural substrate of disrupted empathy in BPD. Our findings on TPJ further show a dissociation between BPD and controls in how empathy correlates with neural activity during social attribution.

Regarding the STG, previous work has shown that patients with BPD exhibited reduced activity in the superior temporal gyrus compared to healthy controls in a cognitive empathy task, while affective empathy was associated with greater insula activity compared to controls (Dziobek et al., 2011). Recent work has also shown a weakened responsiveness of the dACC that may reflect a failure to habituate to negative stimuli and add to the affective instability found in BPD patients (Koenigsberg et al., 2014). With respect to the precuneus, reappraisal strategies correlated with increased activation of the posterior cingulate and precuneus regions in both patients in controls (Koenigsberg et al., 2009). The current study, by contrast, showed that the precuneus was among the regions that discriminated BPD patients from controls, potentially because we took empathy measures into account. Taken together, the neural findings indicate that activity in regions implicated in mentalizing and cognitive reappraisal scaled positively with empathy in BPD patients but negatively in controls. A possible explanation might be that patients with BPD have to activate the mentalizing network to achieve high empathy in general (e.g. self-reported) and when doing the attribution task, while HCC who are more spontaneously (possibly by prior learning) empathic do not need to engage the network as much.

This study had some limitations that should be considered. Due to the cross-sectional study design, we cannot establish a causal link between empathy and social attributions. Next, we relied on self-reports to measure empathy, and deficits in introspection may have introduced more variation in the empathy ratings of BPD patients. Lastly, the sample was relatively small which may increase the chance of what has been termed a Type S error, an error of the estimate being in the wrong direction (Gelman and Carlin, 2014), a limitation that warrants replication of the current findings in larger samples.

# 5. Conclusion

The current study identified a link between empathy and attribution style. The extent to which the emotions of others are shared influences the way their behavior is perceived: in richer and more concrete states instead of abstract, global traits. The disruption of this link in BPD may be related to activity in brain regions implicated in mentalizing when engaged in social attributions.

## Acknowledgments

This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

## Financial Disclosures

All authors declare no conflicts of interest with regard to the current study. Funding was provided by a Klingenstein-Simons Fellowship Award in the Neurosciences to D.S.; and Swiss National Science Foundation grant SNF 161077 to P.H. The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

# References

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Publishing: Arlington.

Bar-Anan Y, Liberman N, Trope Y (2006): The association between psychological distance and construal level: Evidence from an implicit association test. Journal of Experimental Psychology: General 135: 609–622.

Bloom P (2013): The baby in the well. the case against empathy. The New Yorker May 20: 34–38.

Bloom P (2017): Empathy, schmempathy: Response to Zaki. Trends in Cognitive Sciences 21: 60–61.

Brosch T, Schiller D, Mojdehbakhsh R, Uleman JS, Phelps EA (2013): Neural mechanisms underlying the integration of situational information into attribution outcomes. Social Cognitive and Affective Neuroscience 8: 640–6.

Cacioppo JT, Petty RE (1982): The need for cognition. Journal of Personality and Social Psychology 42: 116–131.

Davis, MH (1980): A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology, 10, 85.

Derntl B, Finkelmeyer A, Eickhoff S, Kellermann T, Falkenberg DI, Schneider F, et al. (2010): Multidimensional assessment of empathic abilities: Neural correlates and gender differences. Psychoneuroendocrinology 35: 67–82.

Dinsdale N, Crespi BJ (2013): The borderline empathy paradox: Evidence and conceptual models for empathic enhancements in borderline personality disorder. Journal of Personality Disorders 27: 172–195.

Dziobek I, Preissler S, Grozdanovic Z, Heuser I, Heekeren HR, Roepke S (2011): Neuronal correlates of altered empathy and social cognition in borderline personality disorder. Neuroimage 57: 539–48.

Eklund A, Nichols TE, Knutsson H (2016): Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences 113: 7900–7905.

Fertuck EA, Jekal A, Song I, Wyman B, Morris MC, Wilson ST, et al. (2009): Enhanced 'reading the mind in the eyes' in borderline personality disorder compared to healthy controls. Psychological Medicine 39: 1979–1988.

Flandin G, Friston KJ (2016): Analysis of family-wise error rates in statistical parametric mapping using random field theory. arXiv preprint arXiv:160608199.

Fonagy P, Bateman A (2008): The development of borderline personality disorder-a mentalizing model. Journal of Personality Disorders 22: 4–21.

Forgas JP (1998):On being happy and mistaken: mood effects on the fundamental attribution error. Journal of Personality and Social Psychology 75: 318–331.

Franzen N, Hagenhoff M, Baer N, Schmidt A, Mier D, Sammer G, et al. (2011): Superior 'theory of mind' in borderline personality disorder: An analysis of interaction behavior in a virtual trust game. Psychiatry Research 187: 224 – 233.

Freeman D (2005):Psychological investigation of the structure of paranoia in a non-clinical population. The British Journal of Psychiatry 186: 427–435.

Friston K (2009): Causal modelling and brain connectivity in functional magnetic resonance imaging. PLoS Biology 7: e33.

Gawronski B, Creighton L (2013): Dual process theories. In: Carlston D, editor. The Oxford Handbook of Social Cognition, Oxford University Press: New York, pp 282–312.

Gelman A, Carlin J (2014): Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. Perspectives on Psychological Science 9: 641–651.

Giardina A, Caltagirone C, Oliveri M (2011): Temporo-parietal junction is involved in attribution of hostile intentionality in social interactions: An rTMS study. Neuroscience Letters 495: 150–154.

Gilbert D, Malone P (1995): The correspondence bias. Psychological Bulletin 117: 21–38.

Grinband J, Wager TD, Lindquist M, Ferrera VP, Hirsch J (2008): Detection of time- varying signals in event-related fMRI designs. Neuroimage 43: 509–20.

Gunderson JG (1996): The borderline patient's intolerance of aloneness: Insecure attachments and therapist availability. American Journal of Psychiatry 153: 752–758.

Gunderson JG, Lyons-Ruth K (2008): BPD's interpersonal hypersensitivity phenotype: A gene-environment-developmental model. Journal of Personality Disorders 22: 22–41.

Haas BW, Miller JD (2015): Borderline personality traits and brain activity during emotional perspective taking. Personality Disorders: Theory, Research, and Treatment 6: 315– 320.

Irle E, Lange C, Sachsse U (2005): Reduced size and abnormal asymmetry of parietal cortex in women with borderline personality disorder. Biological Psychiatry 57: 173–182.

Koenigsberg HW, Denny BT, Fan J, Liu X, Guerreri S, Mayson SJ, et al. (2014): The neural correlates of anomalous habituation to negative emotional pictures in borderline and avoidant personality disorder patients. American Journal of Psychiatry 171: 82–90.

Koenigsberg HW, Fan J, Ochsner KN, Liu X, Guise KG, Pizzarello S, et al. (2009): Neural correlates of the use of psychological distancing to regulate responses to negative social cues: a study of patients with borderline personality disorder. Biological Psychiatry 66: 854–63.

Koenigsberg HW, Harvey PD, Mitropoulou V, New AS, Goodman M, Silverman J, et al. (2001):

Are the interpersonal and identity disturbances in the borderline personality disorder criteria linked to the traits of affective instability and impulsivity? Journal of Personality Disorders 15: 358–370.

Krause-Utz, A, Winter, D, Niedtfeld, I, Schmahl, C (2014): The latest neuroimaging findings in Borderline Personality Disorder. Current Psychiatry Reports 16: 438.

Lamm C, Majdanduic J (2015): The role of shared neural activations, mirror neurons, and morality in empathy - critical comment. Neuroscience Research 90: 15–24. □

Lavoie MA, Vistoli D, Sutliff S, Jackson PL, Achim AM (2016): Social representations and contextual adjustments as two distinct components of the theory of mind brain network: Evidence from the Remics task. Cortex 81: 176–191.

Mayer JD, Caruso DR, Salovey P (1999): Emotional intelligence meets traditional standards for an intelligence. Intelligence 27: 267–298.

Mier D, Lis S, Esslinger C, Sauer C, Hagenhoff M, Ulferts J, et al. (2013): Neuronal correlates of social cognition in borderline personality disorder. Social Cognitive and Affective Neuroscience 8: 531–537.

Miller PR, Dasher R, Collins R, Griffiths P, Brown F (2001): Inpatient diagnostic assessments: 1. accuracy of structured vs. unstructured interviews. Psychiatry Research 105: 255–64.

Minzenberg MJ, Poole JH, Vinogradov S (2006): Social-emotion recognition in borderline personality disorder. Comprehensive Psychiatry 47: 468–474.

Nussbaum S, Trope Y, Liberman N (2003): Creeping dispositionism: the temporal dynamics of behavior prediction. Journal of Personality and Social Psychology 84: 485–497.

Pompili M, Girardi P, Ruberto A, Tatarelli R (2005): Suicide in borderline personality disorder: A meta-analysis. Nordic Journal of Psychiatry, 59: 319-324.

Preißler S, Dziobek I, Ritter K, Heekeren HR, Roepke S (2010): Social cognition in borderline personality disorder: Evidence for disturbed recognition of the emotions, thoughts, and intentions of others. Frontiers in Behavioral Neuroscience 4.

R Core Team (2016): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rifkin J (2010): The Empathic Civilization: The Race to Global Consciousness in a World in Crisis. Los Angeles: Jeremy P. Tarcher Inc.

Rim S, Uleman JS, Trope Y (2009): Spontaneous trait inference and construal level theory: Psychological distance increases nonconscious trait thinking. Journal of Experimental Social Psychology 45: 1088–1097.

Roepke S, Vater A, Preißler S, Heekeren HR, Dziobek I (2013): Social cognition in borderline personality disorder. Frontiers in Neuroscience 6.

Ross L (1977): The intuitive psychologist and his shortcomings: Distortions in the attribution process. In: Berkowitz L, editor. Advances in social psychology. Academic Press: New York., New York, NY: Academic Press, pp 173–220.

Saxe R, Kanwisher N (2003): People thinking about thinking people. the role of the temporo-parietal junction in ”theory of mind”. Neuroimage 19: 1835–1842.

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.

Skodol AE, Gunderson JG, Pfohl B, Widiger TA, Livesley W, Siever LJ (2002): The borderline diagnosis I: Psychopathology, comorbidity, and personality structure. Biological Psychiatry 51: 936–950.

Savalia NK, Agres PF, Chan MY, Feczko EJ, Kennedy KM, Wig GS (2017): Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. Human Brain Mapping 38: 472-492.

Stephan E, Liberman N, Trope Y (2011): The effects of time perspective and level of construal on social distance. Journal of Experimental Social Psychology 47: 397–402.

Tomko RL, Trull TJ, Wood PK, Sher KJ (2013): Characteristics of borderline personality disorder in a community sample: Comorbidity, treatment utilization, and general functioning. Journal of Personality Disorders 27: 1–17.

Trope Y (1986): Identification and inferential processes in dispositional attribution. Psychological Review 93: 239–257.

Trope Y, Liberman N (2010): Construal-level theory of psychological distance. Psychological Review 117: 440–463.

Van Overwalle F, Baetens K (2009): Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. NeuroImage 48: 564–584.

| **Characteristics** | **HCC** | **Mean** | **SD** | **BPD** | **Mean** | **SD** | ***t*** | ***P*** |
|---|---|---|---|---|---|---|---|---|
| Males | 7 | | | 7 | | | | |
| Females | 10 | | | 11 | | | | |
| Age | 17 | 37.4 | 12.4 | 18 | 36.2 | 11.1 | -0.3 | 0.78 |
| STAIT | 7 | 26.6 | 4.7 | 16 | 44.6 | 9.5 | 6.1 | < 0.001 |
| STAIS | 7 | 24.9 | 3.8 | 16 | 39.2 | 11.9 | 4.3 | < 0.001 |
| BDI | 9 | 2.3 | 1.5 | 16 | 16.9 | 10.3 | 5.5 | < 0.001 |
| IRI | 15 | -0.9 | 2.5 | 14 | -1 | 3.1 | -0.1 | 0.95 |
| IRI perspective taking | 15 | 2.8 | 0.6 | 14 | 2.2 | 0.7 | -2.5 | 0.02 |
| IRI fantasy | 15 | 1.9 | 0.7 | 14 | 2.1 | 0.4 | 1 | 0.32 |
| IRI empathy | 15 | 2.9 | 0.5 | 14 | 2.4 | 0.5 | -2.5 | 0.02 |
| IRI distress | 15 | 1.3 | 0.6 | 14 | 1.5 | 0.7 | 0.8 | 0.42 |
| Need for cognition | 15 | 67.7 | 12.5 | 14 | 62.2 | 10.6 | -1.3 | 0.21 |
| Empathy | 15 | 3.8 | 0.5 | 14 | 3.2 | 0.6 | -2.5 | 0.02 |
| Empathy suffering | 15 | 4.3 | 0.5 | 14 | 3.7 | 0.9 | -2.2 | 0.04 |
| Empathy positive sharing | 15 | 4 | 0.8 | 14 | 3.3 | 1 | -2.1 | 0.04 |
| Empathy responsive crying | 15 | 3 | 1.2 | 14 | 2.6 | 1.1 | -0.8 | 0.4 |
| Empathy emotional attention | 15 | 4.3 | 0.6 | 14 | 3.3 | 0.9 | -3.5 | < 0.001 |
| Empathy feel for others | 15 | 2.7 | 0.9 | 14 | 2.9 | 0.9 | 0.5 | 0.6 |
| Empathy emotional contagion | 15 | 3.2 | 0.8 | 14 | 2.8 | 0.8 | -1.5 | 0.14 |
| Paranoia | 15 | 20.4 | 4.4 | 14 | 35.4 | 10.9 | 4.8 | < 0.001 |
| Paranoia frequency | 15 | 19.4 | 4.2 | 14 | 34.1 | 12.4 | 4.2 | < 0.001 |
| Paranoia conviction | 15 | 19.5 | 3.8 | 14 | 34.2 | 12.3 | 4.3 | < 0.001 |
| Paranoia distress | 15 | 22.1 | 9.3 | 14 | 37.7 | 15 | 3.3 | < 0.001 |

**Table 1: Sample characteristics**. The number of participants for each characteristic is indicated in the columns for HCC and BPD, respectively. *Abbreviations*: HCC, healthy control cases; BPD, patients with borderline personality disorder; SD, Standard deviation STAI-S/STAI-T, state/trait anxiety subscale of the Spielberger State-Trait Anxiety Inventory; BDI, Beck Depression Inventory; IRI, Interpersonal Reactivity Index; *t*, t-test statistic, *P*, *P*-value.

| **Predictor** | **Estimate** | **SE** | ***t*** | ***P*** |
|---|---|---|---|---|
| *Model:* trait attributions = group + empathy + group * empathy | | | | |
| Intercept | 0.697 | 0.203 | 3.429 | 0.002 |
| Group | 0.865 | 0.35 | 2.475 | 0.02 |
| Empathy | -0.057 | 0.062 | -0.927 | 0.363 |
| Group * Empathy | -0.207 | 0.097 | -2.134 | 0.043 |
| | | | | |
| *Model:* trait attributions = group + paranoia + group * paranoia | | | | |
| Intercept | 0.413 | 0.165 | 2.502 | 0.019 |
| Group | 0.239 | 0.276 | 0.864 | 0.396 |
| Paranoia | 0.003 | 0.004 | 0.627 | 0.536 |
| Group * Paranoia | -0.007 | 0.012 | -0.605 | 0.55 |
| | | | | |
| *Model:* trait attributions = group + need for cognition + group * need for cognition | | | | |
| Intercept | 0.694 | 0.291 | 2.383 | 0.025 |
| Group | -0.059 | 0.39 | -0.152 | 0.88 |
| Need for cognition | -0.003 | 0.005 | -0.633 | 0.533 |
| Group * Need for cognition | 0.002 | 0.006 | 0.321 | 0.751 |

**Table 2: Analyses of covariance of trait attributions**. Estimates indicate the strength of the corresponding predictor. *Abbreviations*: SE, standard error; t, t-statistic; *P*, *P*-value.

| **Region** | **P_cluster** | **k** | **P_peak** | ***t*** | **x** | **y** | **z** |
|---|---|---|---|---|---|---|---|
| Precuneus, left | 0.001 | 455 | 0.004 | 7.06 | -12 | -42 | 48 |
| MFG, right | <0.0001 | 577 | 0.006 | 6.87 | 12 | -20 | 50 |
| TPJ, right | <0.0001 | 1655 | 0.012 | 6.52 | 60 | -28 | 30 |
| STG, left | <0.0001 | 930 | 0.072 | 5.62 | -62 | -26 | 14 |
| dACC, left | 0.004 | 346 | 0.218 | 5.04 | -8 | 4 | 36 |

**Table 3: Neural correlates of group-by-empathy interaction**. Listed are activations that survived a voxel level threshold of P < 0.001 (uncorrected) and a cluster level threshold of P < 0.05 (corrected for family-wise error). *Abbreviations*: MFG, medial frontal gyrus; TPJ, temporoparietal junction; STG, superior temporal gyrus; dACC, dorsal anterior cingulate cortex; P_cluster, cluster *P*-value (corrected for FWE); P_peak, voxel peak *P*-value (corrected for FWE); *t*, voxel-level *t*-statistic; x, y, z, Montreal Neurologic Institute coordinates

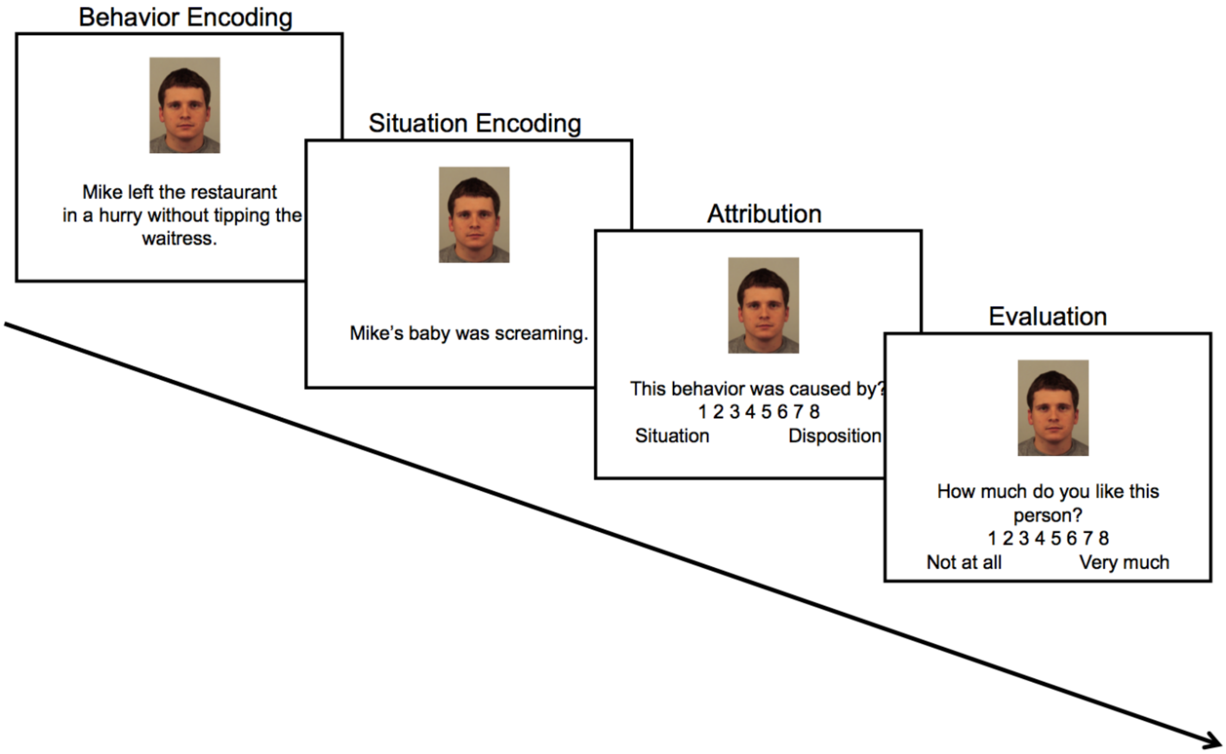
Behavior Encoding
Mike left the restaurant in a hurry without tipping the waitress.
Situation Encoding
Mike's baby was screaming.
Attribution
This behavior was caused by?
1 2 3 4 5 6 7 8
Situation
Disposition
Evaluation
How much do you like this person?
1 2 3 4 5 6 7 8
Not at all
Very much

**Figure 1: Social judgment task**. There were 32 vignettes consisting of a behavioral and a situational segment (6 s each; intertrial interval 2 - 6 s; order counterbalanced). Following each vignette, the participants answered questions about the cause of the behavior and how much they like the person (maximum duration 10.5 s; questions order counterbalanced). One half of the vignettes described a positive behavior, the other half a negative behavior. Notably, each situational segment was constructed in a way that could potentially relativize both the positive and the negative behaviors.

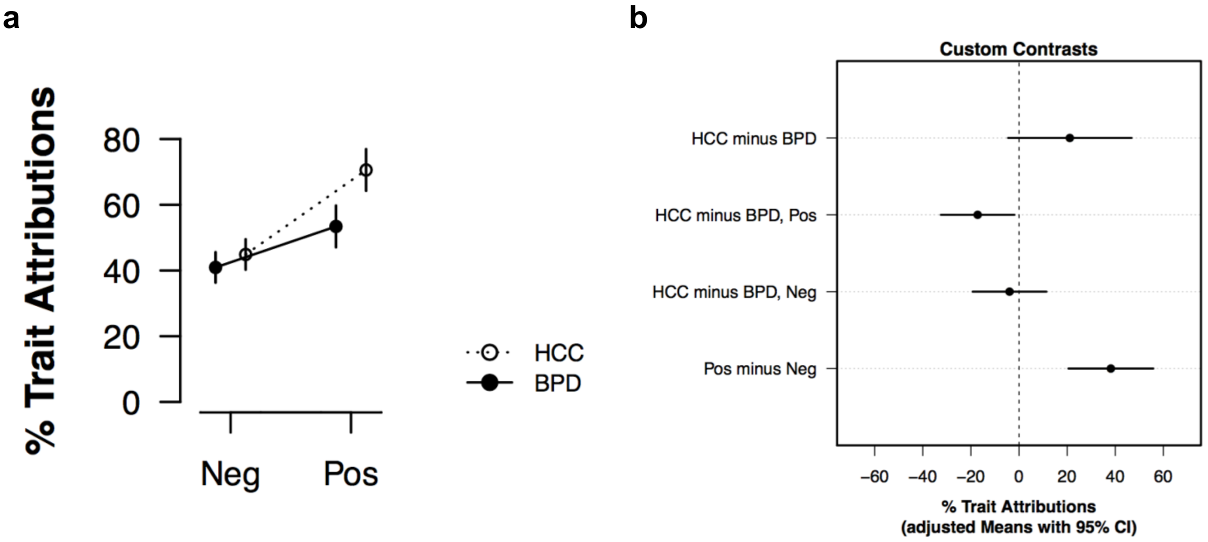

**Figure 2: Healthy controls and BPD patients did not differ in their overall tendency to attribute behavior to traits and made more trait attributions when judging positive compared to negative behavior. a**, Mean percentage of trait attributions by behavior valence. Error bars correspond to standard errors of the mean. **b**, Custom contrasts confirmed the absence of a group difference on trait attribution irrespective of behavior valence and the absence of the interaction of group-by-valence. Note that although the groups did differ when judging positive behavior they did not significantly differ for negative behavior, and neither the interaction nor the main effect of group were significantly different. Both groups made significantly more trait attributions for positive compared to negative behaviors. Adjusted means indicate the mean response for each factor, adjusted for any other variables in the model. Error bars correspond to 95% confidence intervals. Confidence intervals that do not include zero (cross the vertical dashed line) indicate that the corresponding contrast is statistically significant. *Abbreviations*: HCC, healthy control cases; BPD, borderline personality disorder; Neg, negative behavior; Pos, positive behavior; CI, confidence interval.

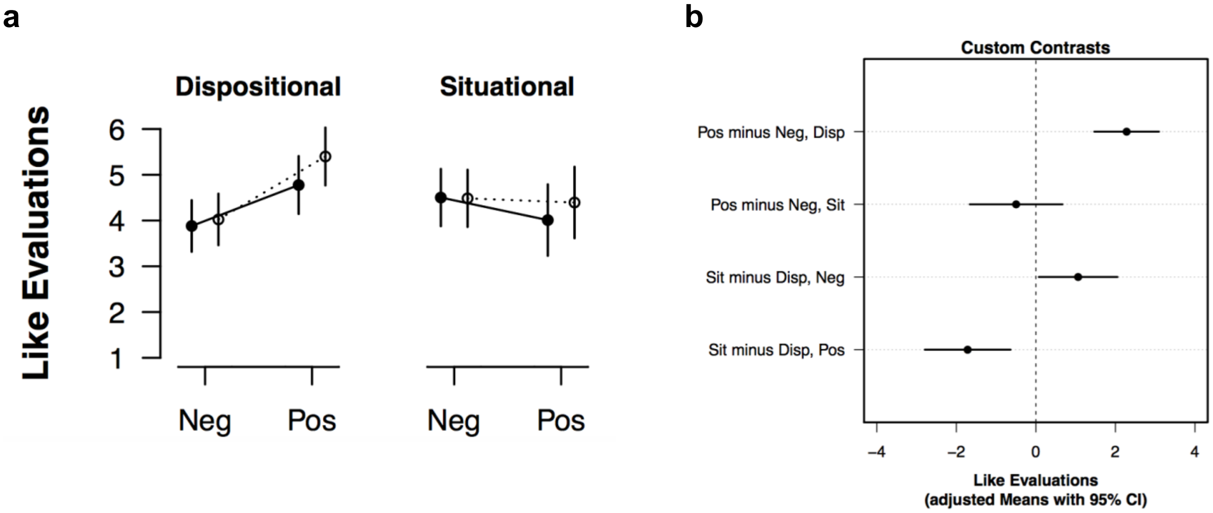

**Figure 3: Healthy controls and BPD patients showed no difference in liking evaluations. a, Mean liking evaluations by behavior valence.** Confirming task validity, liking evaluations were in line with the attribution judgments: participants liked a person more when that person had displayed a positive compared to a negative behavior that they attributed to the personality (dispositional judgment); but participants adjusted their evaluations when they attributed the behavior to the situation (situational judgment). Error bars correspond to standard errors of the mean. **b**, Custom contrasts show that both groups rated positive behavior significantly higher than negative behavior when they had made a dispositional judgment, while this difference was not significant for situational judgments. In addition, ratings for negative behaviors increased and for positive behaviors decreased significantly when a situational compared to a dispositional judgment had been made. Adjusted means indicate the mean response for each factor, adjusted for any other variables in the model. Error bars correspond to 95% confidence intervals. Confidence intervals that do not include zero (cross the vertical dashed line) indicate that the corresponding contrast is statistically significant. *Abbreviations*: Disp, dispositional judgment; Sit, situational judgment; HCC, healthy control cases; BPD, borderline personality disorder; Neg, negative behavior; Pos, positive behavior; CI, confidence interval.

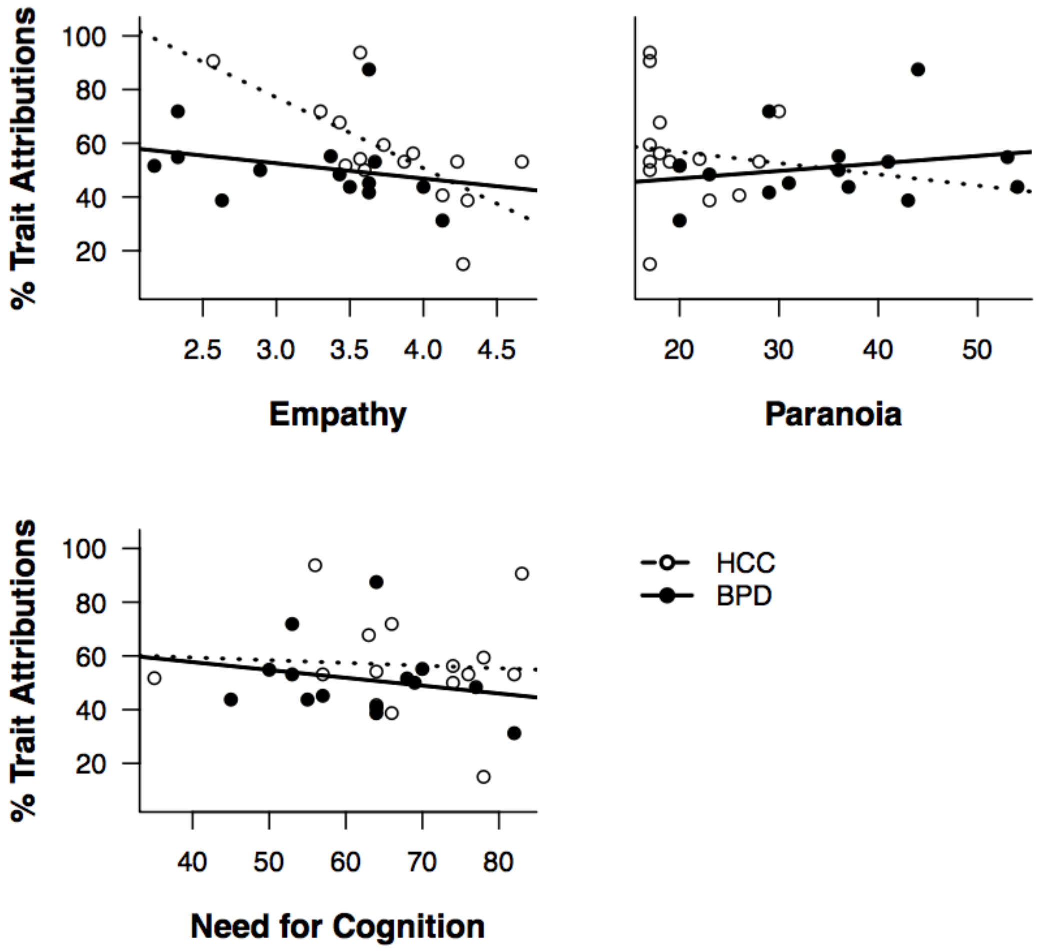

**Figure 4: Regressions of trait attributions and empathy, paranoia, and need for cognition scores in healthy controls and BPD patients.** Empathy was significantly less informative in predicting percent trait attributions in patients compared to controls, no such relationship was found for paranoia and need for cognition. *Abbreviations*: HCC, healthy control cases; BPD, borderline personality disorder.

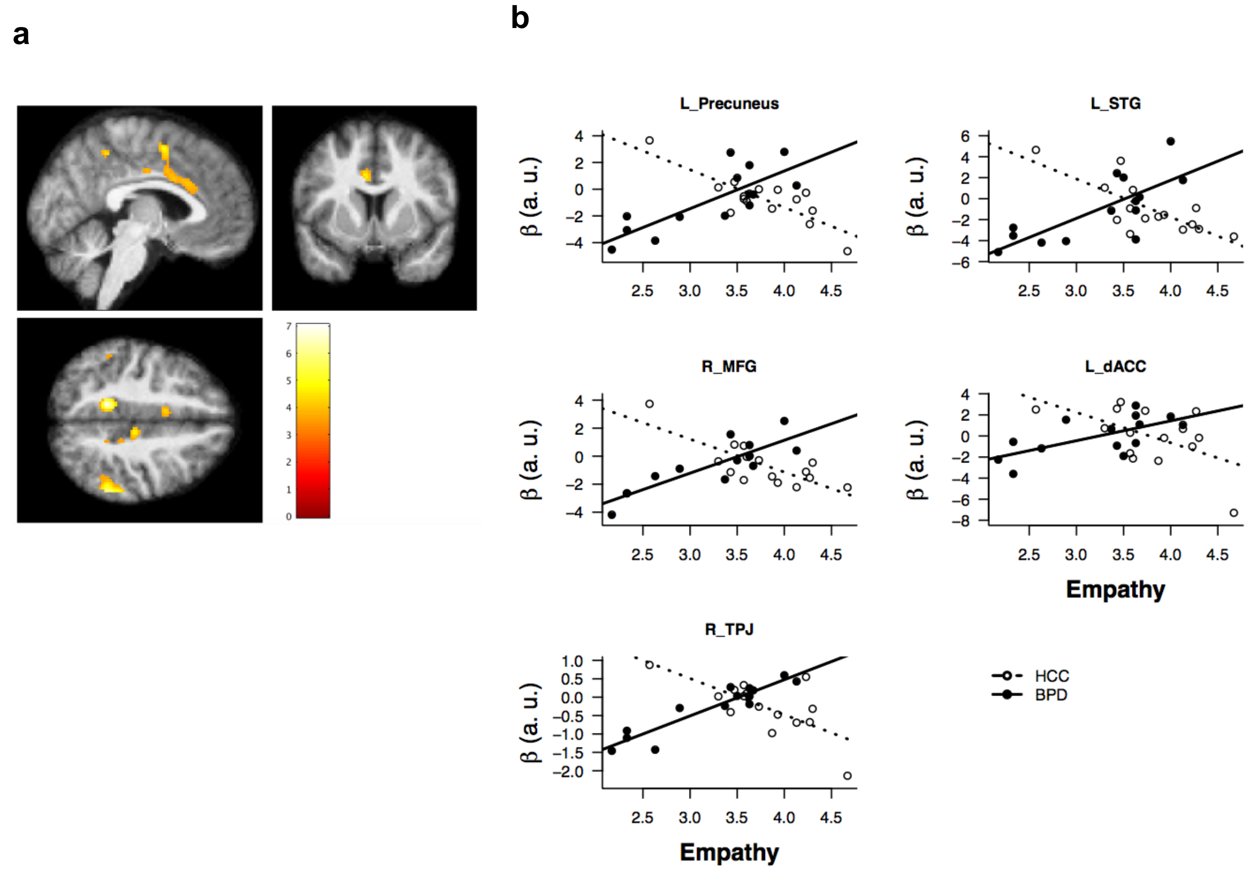

**Figure 5: Neural correlates of social attributions in healthy controls and BPD patients. a, Whole brain analysis of group-by-empathy interaction.** When participants judged the behavior during the attribution phase, several brain regions implicated in mentalizing, robustly distinguished patients from controls in their relationship with empathy. Slices are shown at a voxel level threshold of $P < 0.001$ and a cluster-level threshold of $P < 0.05$, corrected for family-wise error. **b, Regression plots of empathy against social attribution beta estimates in healthy controls and BPD patients.** All regions show a similar pattern of neural activity that scales positively with empathy in patients but negatively in healthy controls. These plots do not provide an independent measure of effect size, but are used here to visualize the direction of the effects that drove the interaction, i.e., a negative relationship between empathy and neural activity in controls compared to a positive relationship in patients. *Abbreviations*: L, left; R, right; MFG, medial frontal gyrus; TPJ, temporoparietal junction; STG, superior temporal gyrus; dACC, dorsal anterior cingulate cortex; HCC, healthy control cases; BPD, borderline personality disorder; a. u., arbitrary units.

## Financial Disclosures

All authors declare no conflicts of interest with regard to the current study. Funding was provided by a Klingenstein-Simons Fellowship Award in the Neurosciences to D.S.; and Swiss National Science Foundation grant SNF 161077 to P.H. The funding source had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.