Title: "Modulation of Trust in Borderline Personality Disorder by Script-Based Imaginal Exposure to Betrayal"

Author List: Shapiro-Thompson R 1, Shah TV 2, Yi C 3, Jackson N 4, Trujillo Diaz D 5, and Fineberg SK 1.

Affiliations:
1 Yale University Department of Psychiatry
2 University of Pittsburgh Department of Psychiatry
3 Yale College
4 Meharry Medical College
5 University of Utah Department of Psychiatry

Figures: 3
Tables: 4

Abstract

Interpersonal and trust-related difficulties are central features of Borderline Personality Disorder (BPD). In this study, we applied script-driven betrayal imagery to evoke mistrustful behavior in a social reinforcement learning task. We compared this approach to the standard confederate paradigm in twenty-one BPD and twenty healthy control (HC) participants. The script-driven imagery evoked transient negative affect and decreased trusting behavior in both groups. Across conditions, we also replicated previously reported between-group differences in affect and task behavior, results that support the validity of script-driven imagery as an alternative social task stimulus. This approach is appealing for eliminating deception, scaling easily, and evoking disorder-specific states of social difficulty.

Introduction

Knowing who to trust and when to change your mind is critical to navigating the social world. For people with Borderline Personality Disorder (BPD), it is difficult to assess trustworthiness (Fertuck, Grinband, & Stanley, 2013; Miano, Fertuck, Arntz, & Stanley, 2013) and to communicate one's own trustworthiness (Dammann et al., 2020). These deeply distressing difficulties are amplified under stressful conditions and in the context of interpersonal problems (Gunderson, Herpertz, Skodol, Torgersen, & Zanarini, 2018).

The significance of interpersonal difficulties is further underlined by the fact that each of the evidence-based psychotherapies for BPD focuses on improving social functioning (Choi-Kain, Finch, Masland, Jenkins, & Unruh, 2017; Gunderson, Fruzzetti, Unruh, & Choi-Kain, 2018). Despite this, interpersonal problems continue to be some of the most persistent symptoms of BPD (Gunderson et al., 2011).

Historically, research in social domains in BPD has heavily leaned toward non-interactive paradigms. Negative social biases have been well established with static stimuli such as facial emotion perception paradigms (e.g., (Fertuck et al., 2019; Fertuck et al., 2013; Miano et al., 2013; Unoka, Fogd, Fuzy, & Csukly, 2011), while neuroimaging research has demonstrated that people with BPD have amygdala hyperactivity and hyper-reactivity to social stimuli, (Schulze, Schmahl, & Niedtfeld, 2016). In the last fifteen years, more interactive approaches have added to our understanding of the processes associated with social rejection (Domsalla et al., 2014; Liebke et al., 2018) and real-time social conflict (King-Casas et al., 2008; Polgar, Fogd, Unoka, Siraly, & Csukly, 2014; Unoka, Seres, Aspan, Bodi, & Keri, 2009). To improve the efficacy of treatments overall and the matching of individual patients to specific treatments, we need to

continue work in interactive paradigms and to measure mechanisms of social functioning <u>in the</u> <u>states that matter</u> (when people with BPD are in the throes of social problems).

In this study, we used a patient-derived betrayal narrative paired with an imaginal autobiographical exposure as a social stimulus to evoke BPD-relevant social experience prior to engagement with a social reinforcement learning task. While script-driven imagery is well-established as a method to evoke disorder-specific experiences in psychiatric patients, we are not aware of previous studies using this approach to evoke relevant symptom states for the examination of trust or reinforcement learning behavior.

Script-driven imagery has been demonstrated to be tolerable and feasible for patients in a wide range of substance use and trauma research studies. Personalized scripts reliably evoke the state-dependent outcomes of interest (e.g. cravings (Ashare et al., 2012; Li, Kosten, & Sinha, 2005; Li & Sinha, 2006), depressed affect (Robins, 1988) and trauma-related symptoms (reviewed in (Bujarski, Mischel, Dutton, Steele, & Cisler, 2015)). Script-driven imagery has also been used in BPD to evoke anger and aggressive feelings (Krauch et al., 2018; Ueltzhoffer, Herpertz, Krauch, Schmahl, & Bertsch, 2019), to evoke trauma-related dissociation (Bichescu-Burian, Grieb, Steinert, Uhlmann, & Steyer, 2018), to understand the relationship of dissociation and pain (Chung, Hensel, Schmidinger, Bekrater-Bodmann, & Flor, 2020), and to identify physiologic (Gratz, Richmond, Dixon-Gordon, Chapman, & Tull, 2019) and neural (Kraus et al., 2010) correlates of non-suicidal self-injury. Script-driven imagery with rejection and abandonment scripts (similar to the procedure we used here) has been shown specifically to reduce ability to propose solutions to social problems (Dixon-Gordon, Chapman, Lovasz, & Walters, 2011) and to potentiate physiologic threat avoidance (eyeblink startle response) more than a non-social negative script (Limberg, Barnow, Freyberger, & Hamm, 2011).

We compared responses to our social betrayal script-driven imagery procedure (script condition) to responses to our standard pre-task social stimulus of meeting a confederate (confederate condition) in two groups: participants with BPD and participants without psychopathology (healthy control, HC). We measured both general affective state and trust behavior in the reinforcement learning task. We think this novel application of script-driven imagery offers an exciting way forward in social learning research: it transiently evokes personal history of interactive social situations, is easily scalable and portable to other study settings, and eliminates the need for deception in this vulnerable patient group.

We tested the following *a priori* hypotheses:

H1 Impacts of Betrayal Script on Affect

a) BPD participants will have lower positive affect and higher negative affect compared to healthy controls overall.

b) Betrayal scripts will have an equal or greater impact on affect compared to meeting the confederate.

c) The betrayal script will have an equal or greater impact on affect in BPD participants than in HC participants.

H2: Impacts of Betrayal Script on Task Behavior

a) Initial cooperation will be less in BPD than HC participants. Initial cooperation will be equal or reduced in the script condition as compared to the confederate condition.

b) Advice-taking overall will be less in BPD than HC participants. Advice-taking will be equal or reduced in the script condition as compared to the confederate condition.

c) The impact of a recent betrayal on advice-taking will be greater in BPD than HC participants. The impact will be equal or greater in the script condition.

d) Impact of social reward probability on choice will be greater in BPD than HC participants. The impact will be equal or greater in the script condition.

H3. Impact of Individual Differences on Task Behavior

a) Within the BPD group, BPD symptom burden will predict less advice-taking.

b) Within the BPD group, the impact of BPD symptom burden on advice-taking will be equal or greater in the script compared to confederate condition.

Methods

Ethics

This protocol was written and conducted in accordance with the Declaration of Helsinki and was approved by the Yale Institutional Review Board (protocol 2000026215). This project involved deception, and all participants who experienced the deception were debriefed at the end of the study and given the opportunity to talk with study staff about their experience and their emotions after hearing about the deception.

Participants

Adults aged 18-60 were recruited from the community. We assessed BPD symptoms in a semi-structured interview using the Diagnostic Interview for Personality Disorders (DIPD) BPD questions (Zanarini, Frankenburg, Chauncey, & Gunderson, 1987) and included people in the BPD group if they met ≥ 5/9 criteria, or the healthy control (HC) group if they met 1 or fewer criteria. We confirmed group assignment by checking self-reported BPD symptoms on the BSL-

23 scale; this resulted in removal of one healthy control participant with markedly discrepant interview versus self-reported symptoms (not included in any analysis reported here). Exclusion criteria for both groups were lifetime diagnosis of primary psychotic disorder, neurologic illness, autism, lifetime history of learning disability, color blindness, and current problems interfering with computer use. Healthy controls were excluded for any lifetime psychiatric diagnosis or treatment, including psychotherapy.

We report participant demographics in **Table 1.** Overall, our sample was predominantly white, female, and educated. The BPD and HC groups did not differ significantly in age, gender, racial groups, or education levels. For age, Mann-Whitney test revealed $U = 163$, $p = 0.225$. Fischer's exact tests were not significant for gender ($p = 1.00$), sex at birth ($p = 0.410$), race ($p = 1.00$), or ethnicity ($p = 1.00$). We analyzed education data by comparing participants with at least a Bachelor's degree (BA) and those with less than a BA (high school, some college, or Associate's degree). A Chi-squared test revealed no significant differences between groups ($x^2(1) = 1.39$, $p = 0.238$).

We also examined self-reported symptom scores for BPD and depression (**Table 2**). We used the twenty-three item Borderline Symptom List (BSL-23) self-report scale to assess BPD symptom severity (Bohus et al., 2009). According to norms recently established by Kleindienst et al. for this scale, our BPD sample ranged from moderate to very high severity, with the mean in the high-moderate range at 1.58 (Kleindienst, Jungkunz, & Bohus, 2020). Participants in the HC group had BSL-23 scores ranging from none or low to mild, with the mean in the none or low range. Depression scores on the Beck Depression Inventory (BDI-II) self-report scale were minimal in the HC group and moderate to severe in the BPD group (Smarr & Keefer, 2011).

Mann-Whitney tests revealed significant between-group differences with large effect sizes for both scales (results in **Table 2**).

Procedure

Research visits were conducted by video conference on Zoom. The first visit began with a consent process; all participants gave written informed consent before any other visit procedures. To determine eligibility and group assignment, we conducted a semi-structured interview using the BPD section of the Diagnostic Interview for Personality Disorders (DIPD) (Zanarini et al., 1987) and in-house questions to check the above-described eligibility criteria. Participants then completed a battery of self-report measures selected for their common use in the field with established validity. In this manuscript, we report on the short form of the Borderline Symptom List (BSL-23, for intensity of recent BPD symptoms, (Bohus et al., 2009)) and the Beck Depression Inventory (BDI-II, for depressive symptoms (Beck, Steer, Ball, & Ranieri, 1996)). Then participants were invited to complete task visits on two subsequent days (**Figure 1**). On the first task visit they were introduced to a remote confederate (confederate condition) before playing the task. On the second visit, they listened to a standardized script about a betrayal (script condition.) Participants completed the Positive and Negative Affect Scale (PANAS, (Watson, Clark, & Tellegen, 1988)) to measure current moment (state) affect on both task days at three timepoints: before the social stimulus, after the social stimulus, and after the reinforcement learning task.

Betrayal Script Development

We developed auditory stimuli consisting of standardized narratives of social betrayal following the script-driven imagery method of Sinha and Tuit (Sinha & Tuit, 2012). The narratives were derived from stories of personal betrayal that we collected from a separate sample of outpatients with BPD.

We then standardized these narratives into one-minute audio scripts. We de-identified the stories and generalized some details to better relate to a broader audience of potential participants. The final scripts were recorded in the standard second-person format, emphasizing physiological sensations (e.g., "Your stomach is in a knot. You hear the door open, and you see your ex."). The script used in this study describes an ex-partner moving out of a previously shared apartment under surprising and hurtful circumstances (full script in **Supplemental Materials**).

Trust Behavior

On both task days, participants completed The Social Valuation Task (SVT) (Behrens, Hunt, Woolrich, & Rushworth, 2008; Fineberg et al., 2018; Henco et al., 2020). The SVT is a reinforcement learning task with cues that are non-social (color) and social (partner advice). Here, we examined how the pre-task social stimulus (confederate or script) impacted participants' trust-related behavior.

We examined several measures of trust in the task behavior:

1) *Initial cooperation* is if the participant took the advice or not on the first trial. We tested this outcome in H2a.

2) *Frequency of advice-taking* is the fraction of trials when the participant took the advice. We tested this outcome in H2b.

3) *Recent betrayal* is a trial when the advice was not rewarded on the previous trial. In H2c, we tested the impact of recent betrayal on advice-taking.

4) *Social reward probability* is the current trial estimate of reward probability for the social cue (Behrens et al., 2008). In H2d, we tested the impact of social reward probability on advice-taking.

Confederate Condition

In the confederate condition, participant and confederate met together with the experimenter to complete a brief icebreaker, task instructions, and task demonstration. This meeting served to establish rapport between participant and confederate and to reinforce the fiction of the confederate as second task player. Task confederates were 20-30 year old Asian and White women trained for consistent performance during the meeting and task demonstration (each one made the same icebreaker responses and same demonstration task choices). After the task demonstrations, confederates left the Zoom meeting, supposedly to set up their own version of the task. The participants played the SVT with the social cues supposedly coming from the partner they just met.

Script Condition

To avoid revealing the deception before the confederate condition, the script condition was always presented on day two, approximately one week after the confederate condition. In the script condition, participants were told that instead of meeting a game partner, they would listen to a story. We asked participants to close their eyes and listen to the betrayal script and to

"imagine this situation as if it were happening right now." We confirmed that they heard the script. We then asked them to recall a situation from their own life that elicited similar feelings. They wrote down a brief description of the autobiographical situation and the name of the person involved. We then started the task, asking them to imagine playing with that person at the time of that past event.

Data Analysis

We conducted statistical analyses in R (Team, 2014). We defined significance as $p < 0.05$ and trend-level significance as $p < 0.1$. For each of the significant results discussed in the paper, we report the test statistic, p-value, and effect size. Effect sizes are reported as Cohen's d for t-tests (conventional interpretation: small = 0.1, medium = 0.3, large = 0.5, though others have suggested smaller values may be more appropriate in individual differences research e.g., (Gignac & Szodorai, 2016) and as generalized eta squared for repeated-measures ANOVAs (conventional interpretation: small ($\eta^2 = 0.01$), medium ($\eta^2 = 0.06$), large ($\eta^2 = 0.14$).

For t-tests, we tested the assumption of normality, and if data were non-normal, we proceeded with a Mann-Whitney non-parametric test. For categorical variables, including gender, sex assigned at birth, race, and ethnicity, we used Fisher's exact tests as the data contained frequencies of less than five and Chi-square test for frequencies of at least five.

We analyzed the affect reports (PANAS-positive and PANAS-negative scores) with respect to group, timepoint, and condition using repeated-measures ANOVAs (**Figure 2**). We also found that five observations from three participants were outliers, and so we tested our models with and without those three participants. We also found that while the positive PANAS data were approximately symmetrically distributed (skewness = 0.39), the negative PANAS data

was highly skewed to the right (skewness = 1.7). A log transformation on the PANAS-negative

data decreased the skew from extreme to moderate (skewness = 0.87, and we ran the ANOVA

models with and without the log transformation. We report the results from the untransformed

data here because neither the removal of outliers nor the log transformation impacted the results.

This is consistent with literature describing ANOVA being robust to normality violation (Blanca,

Alarcon, Arnau, Bono, & Bendayan, 2017).

We analyzed trust-related task behavior with respect to group, timepoint, and condition

(**Figure 3**). We used repeated-measures ANOVA to test main effects of group, condition, and

timepoint and the two- and three-way interactions on initial cooperation (H2a) and frequency of

advice-taking (H1b). We used logistic regressions to test the impacts of recent betrayal (H2c),

social reward probability (H2d), and BPD participant symptom burden (H3) on advice-taking.

See models and results in **Table 3** and **Table 4**.

Results

H1 Impacts of Betrayal Script on Affect

H1a) BPD participants will have lower positive affect and higher negative affect compared to

healthy controls overall.

As hypothesized, we found main effects of group on both positive and negative affect in

repeated-measures ANOVAs (for positive affect, $F(1, 231) = 11.59$, $p < 0.001$, $\eta^2 = 0.048$; for

negative affect, $F(1, 231) = 20.98$, $p < 0.001$, $\eta^2 = 0.083$, **Figure 2**). Overall, BPD participants

had lower positive affect and higher negative affect than HC participants.

H1b) Betrayal scripts will have an equal or greater impact on affect compared to meeting the confederate.

For negative affect, we found a main effect of timepoint (F(2, 231) = 12.14, < 0.001, η^2 = 0.095). Participants experienced heightened negative affect after the social stimulus compared to baseline (t(2, 137) = 4.05, p < 0.001, d = 0.633). This heightened negative affect then decreased after the task (t(2, 146) = -3.35, p = 0.001, d = 0.527) to a level not significantly different than baseline (t(2, 156) = 0.71, p = 0.47).

As hypothesized, we did find that negative affect increased more after hearing the script than after meeting the confederate. A timepoint by condition interaction was significant (F(2, 231) = 4.12, p = 0.018, η^2 = 0.0340), driven by differences in the post-stimulus timepoint (t(2, 72.4) = 3.20, p = 0.002, d = 0.704). In other words, betrayal scripts had a greater impact on negative affect than meeting the confederate (**Figure 2**).

We found no significant impact of timepoint or condition by timepoint interaction on positive affect.

H1c) The betrayal script will have an equal or greater impact on affect in BPD participants than in HC participants.

When we examined the impacts of the betrayal script on BPD and HC participants, we found no significant group by condition (2 x 2) or group by condition by timepoint (2 x 2 x 3) interaction for either positive or negative affect. While BPD participants had less positive affect and more negative affect overall, the two groups responded similarly to the social stimuli.

H2: Impacts of Betrayal Script on Task Behavior

H2a) Initial cooperation will be less in BPD than HC participants. Initial cooperation will be equal or reduced in the script condition as compared to the confederate condition.

In a repeated-measures ANOVA, we found a trend toward lower initial cooperation in BPD compared to healthy participants across conditions ($F(1, 78) = 2.72$, $p = 0.083$, $\eta^2 = 0.038$) (**Figure 3A**). As for the impact of condition, we found a significant main effect on initial cooperation ($F(1, 78) = 5.76$, $p = 0.019$, $\eta^2 = 0.069$). In the script condition, both groups were less likely to take advice on the first trial. The group by condition interaction was not significant. This result suggests that the script led both BPD and control participants to cooperate less, and that the impact of the script did not differ by group.

H2b) Advice-taking overall will be less in BPD than HC participants. Advice-taking will be equal or reduced in the script condition as compared to the confederate condition.

While in H2a we found a trend-level impact of group on initial cooperation, here we did not find the hypothesized main effect of group on advice-taking. However, we did find the hypothesized main effect of condition with a small effect size ($F(1, 23737) = 49.837$, $p < 0.001$, $\eta^2 = 2.0$ x 10^{-3}), such that participants across both groups were less likely to follow advice in the script condition than in the confederate condition (**Figure 3B**). We also found a trend-level interaction between group and condition with small effect size ($F(1, 23737) = 3.119$, p $= 0.0770$, $\eta^2 = 1.31$ x 10^{-4}) suggesting that scripts may have had a slightly differential impact on the group task behavior. As shown in **Figure 3B**, BPD participants were slightly more likely than HCs to follow advice in the remote confederate condition ($t(2, 11855) = 1.80$, $p = 0.0712$, $d = 0.03$), though the effect size was quite small. In the generalized scripting condition, we found no significant difference in overall advice-taking by group.

H2c) The impact of a recent betrayal on advice-taking will be greater in BPD than HC participants. The impact will be equal or greater in the script condition.

In a logistic regression, we evaluated the impact of a betrayal in the previous trial on current trial advice-taking. Consistent with previous work, we found that betrayal in the previous trial was a strong negative predictor of advice-taking in the current trial (**Table 3**). We also found an interaction between recent betrayal and group whereby betrayal more strongly reduced advice-taking for BPD than HC participants. As hypothesized, we found an interaction between recent betrayal and condition: recent betrayal reduced advice-taking more in script than the confederate condition Notably, we did not find an interaction between recent betrayal, condition, and group, suggesting that the script condition did not have a differential impact on how the groups responded to a recent betrayal.

H2d) Impact of social reward probability on choice will be greater in BPD than HC participants. The impact will be equal or greater in the script condition.

Consistent with previous work (Fineberg et al., 2018), we found that BPD participants were more responsive than healthy controls to the current probability of correct advice (**Table** 3). As hypothesized, all participants were more responsive to the current probability of correct advice in the script than the confederate condition. Here, we did not find a probability by group by condition interaction. This means that the two groups weighted the probability of advice to a similar increased degree in the script condition compared to the confederate condition.

H3. Impact of Individual Differences on Task Behavior

H3a) Within the BPD group, BPD symptom burden will predict less advice-taking.

Within the BPD sample, we conducted a logistic regression comparing the impact of BSL-23 score on advice-taking by condition. We found a significant main effect of BSL-23 score such that, for BPD participants, higher BSL-23 scores predicted less advice-taking in both conditions (**Table 4**).

H3b) Within the BPD group, the impact of BPD symptom burden on advice-taking will be equal or greater in the script compared to confederate condition.

We also found a significant interaction between condition and BSL-23 score (**Table 4**), such that in the script condition, a higher BSL-23 score predicted even less advice-taking than in the confederate condition.

Discussion

Validation of the script procedure as an evocative stimulus for both BPD and HC participants

In this work, we set out to test trust-related behavior following either a patient-derived narrative stimulus or a standard confederate paradigm. We found multiple results that confirm our hypothesis that the narrative stimulus evokes a state of negative affect and mistrust in both BPD and HC participants.

We found that the script evoked negative affect to a greater degree than the confederate for participants in both groups. This is not surprising: the script procedure is both more explicitly negative (a betrayal story instead of meeting a pleasant confederate) and more explicitly personal (participants are asked to think of the individual from an autobiographical betrayal to hold in

mind as the game partner in the script condition). This is consistent with results of previous studies using scripts as stimuli in BPD (Dixon-Gordon et al., 2011; Gratz et al., 2019; Krauch et al., 2018; Robins, 1988; Ueltzhoffer et al., 2019). Though participants with BPD reported more negative affect throughout the process on both test days, it appears that the script procedure was evocative enough to impact on both people with high and low social symptomatology.

Behavior in this task replicates previous social exchange experiments in BPD. King-Casas et al. reported decreased initial cooperation in BPD versus HC participants in their 2008 examination of motivated social exchange in a 10-round Trust Game. Here we also observed decreased initial cooperation in BPD compared to HC participants in both conditions. In our previous work with the SVT, we found that recent experience (wrong advice on the previous trial) and learned probability (current estimate of social reward likelihood) had significant impacts on choice, and those impacts were exaggerated in BPD versus HC participants (Fineberg et al., 2018). Here we replicated that result across conditions.

Given our interest in evoking the states most relevant to social difficulties in BPD, we were particularly interested in whether the script was impactful on trust-related behaviors. Indeed, we found that the script condition had a negative impact on both initial cooperation and overall advice-taking. Differences by overall advice-taking by group and condition were quite subtle; future work will benefit from modelling that takes advantage of trial-wise analysis and hierarchical beliefs as has been done elsewhere with this task (Henco et al., 2020; Sevgi, Diaconescu, Tittgemeyer, & Schilbach, 2015). Using our regression approach, we did observe stronger responses to social betrayal and social reward probability in the script condition as compared to the confederate condition. For BPD participants, current BPD symptom burden

(BSL-23 score) also predicted less advice-taking across conditions and even less advice-taking in the script than in the confederate condition.

Taken together, these data suggest that the script stimulus evokes both affective and behavioral responses consistent with the state of interpersonal difficulty we would like to better understand.

Advantages of the script procedure over other social stimuli

We think this approach has several advantages over other common social stimuli with regards to logistics, ethics, and ecologic validity:

1) The script-based procedure is easy to use. It takes fewer than five minutes total to play the audio file, verify that it was heard and understood, ask the participant to recall an autobiographical event that evoked similar feelings, collect information about the recalled event, and to instruct the participant to hold that event and person in mind for the next steps. The script is easily used in any setting where playing audio is possible, requiring little equipment and no specialized software. In comparison, introducing a confederate introduces logistical complications for scheduling and experimental biases (see below). Ethics boards also have varying views on deception; removing deception can facilitate collaboration and replication and allow for repeated-measures study designs.

2) Though the ethics of work with deception are complicated and empirical work does not clearly suggest participant harm from deception (Boynton, Portnoy, & Johnson, 2013), we would tend to prefer experimental designs that promote positive interactions and grow trust between researchers and research participants. We see this as particularly important for participants who have trust-related symptoms.

3) Though it can be laboratory-based and does not require knowledge of the individual participant's history or relationships, this script procedure explicitly evokes autobiographical social interaction and associated thoughts and feelings. This is because the script procedure presents a story in the second-person ("you felt…") with physiologic detail, and then asks the participant to focus on a similar personal experience. Additionally, this approach side-steps some of the bias that can be introduced by confederate and avatar-based social stimuli, e.g. matches and mis-matches on a host of individual features such as gender, race, and age.

Ecological Validity of Social Interactive Stimuli

Social neuroscience research has often employed static non-interactive stimuli. Leonard Schilbach has called for a shift toward "second person neuroscience" in which social cognition experiments would have participants both emotionally engaged and in interaction with a social partner (Schilbach et al., 2013). This approach tacks toward greater ecologic validity by placing the participant in a closer, related position to the social partner, seeing the partner as a "you" more than a dispassionate "them."

More than requiring a shift in experimental design away from participant as passive observer, this call also asks us to think about how participants perceive social partners in our paradigms. It is now well-documented that participant responses differ in tasks when they believe that a game partner is a human versus a computer. This has been studied in joint attention tasks (Caruana & McArthur, 2019), where partners thought to be human evoked more emotional arousal (Zhang, Dumas, Kelso, & Tognoli, 2016), slower response time (Caruana, de Lissa, & McArthur, 2017), and increased self-reported feeling of the interaction being "natural" (Caruana & McArthur, 2019). In social exchange settings, participants were quicker to stop accepting a

human versus computer partner's bad advice (Goodyear et al., 2016), and more willing to cooperate with a human versus computer partner (McCabe, Houser, Ryan, Smith, & Trouard, 2001). Neuroimaging studies have also identified differences between playing with a perceived human versus computer (Gallagher, Jack, Roepstorff, & Frith, 2002; Pfeiffer et al., 2014; Rauchbauer et al., 2019; Rice & Redcay, 2016; Schindler & Kissler, 2016).

In this study, we found that this script-driven imagery procedure was feasible and tolerable for both BPD and HC participants, and that it had strong impacts on affect and trust behavior. Despite this being a reductionist paradigm that does not rely on facial expression, gesture, proximity, speech or other elements of a live interaction such as real-time mutual engagement, the script method incorporates several emotional and interactive elements called for in Second Person Neuroscience. Specifically, this imaginal exposure can bring real-world relational difficulty and its attendant emotions into the lab.

Limitations & Future Directions

Our study sample is predominantly female and white, though we know that BPD is common in people of all genders, racial groups, and ethnicities. This demographic skew is common to BPD research projects, but we aim in future work to recruit a more representative sample of participants. We acknowledge that mistrust of research programs in BIPOC communities is based on a long history of scientific disinterest and unethical practice; we see it as our responsibility to collaborate with community leaders and potential participants to increase the extent to which our studies occur in a safe and inviting place for everyone.

Our study also used confederates who were all female and White or Asian. This could have introduced in- and out-group effects with our participants and impacted upon trust. These

biases may have been lessened in the script condition and may be avoided in future work by taking script-driven imagery approaches to social tasks.

Though we demonstrated feasibility and tolerability of a script-driven imagery approach to this social task, our study design leaves open questions about potential order effects. All participants here saw the script condition after the confederate condition because the confederate condition depends on the idea that the game works when there is another player. We did not want to risk prematurely revealing the deception by suggesting that the partner could be imaginary (as they are in the script condition). Order effects could have decreased the impact of social cue (habituation) or by contrast increased it (sensitization); there is some evidence of sensitization to negative social cues in BPD (Denny et al., 2018; Koenigsberg et al., 2014). In future work, it will be interesting to develop and test scripts that differ in their levels of intensity, their valence, and their social relevance. Comparing participant response to scripts of differing valences along with a counterbalanced design would allow us to better control for these potential non-associative learning effects.

This study design also does not allow us to disentangle the impact of the script alone from the evocation of autobiographical betrayal. In future work, it could be interesting to more closely examine this by measuring affect after the script but before the autobiographical component is evoked, or to test the relative impacts of script-only and script plus imaginal autobiographical exposure.

Conclusion

In this study, we demonstrated that the script-based social stimulus represents an appealing alternative method to transiently evoke the distressed mistrustful social states most

relevant to suffering in Borderline Personality Disorder. This approach can help to facilitate mechanistic work and the inclusion of social-problem relevant mechanistic outcome measures in treatment studies. Social problems are a painful part of many mental illnesses, and we expect that the script approach may be useful to investigate social learning in a wide range of different people.

Author Statement

References Cited:

Allen, L., Brand, A., Scott, J., Altman, M., & Hlava, M. (2014). Credit where credit is due. *Nature, 508*(7496), 312-313.

Ashare, R. L., Sinha, R., Lampert, R., Weinberger, A. H., Anderson, G. M., Lavery, M. E., . . . McKee, S. A. (2012). Blunted vagal reactivity predicts stress-precipitated tobacco smoking. *Psychopharmacology (Berl), 220*(2), 259-268. doi: 10.1007/s00213-011-2473-3

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *J Pers Assess, 67*(3), 588-597. doi: 10.1207/s15327752jpa6703_13

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature, 456*(7219), 245-249. doi: 10.1038/nature07538

Bichescu-Burian, D. M., Grieb, B., Steinert, T., Uhlmann, C., & Steyer, J. (2018). Use of a Psychophysiological Script-driven Imagery Experiment to Study Trauma-related Dissociation in Borderline Personality Disorder. *J Vis Exp*(133). doi: 10.3791/56111

Blanca, M. J., Alarcon, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29*(4), 552-557. doi: 10.7334/psicothema2016.383

Bohus, M., Kleindienst, N., Limberger, M. F., Stieglitz, R. D., Domsalla, M., Chapman, A. L., . . . Wolf, M. (2009). The short version of the Borderline Symptom List (BSL-23): development and initial data on psychometric properties. *Psychopathology, 42*(1), 32-39. doi: 10.1159/000173701

Boynton, M. H., Portnoy, D. B., & Johnson, B. T. (2013). Exploring the ethics and psychological impact of deception in psychological research. *IRB, 35*(2), 7-13.

Bujarski, S. J., Mischel, E., Dutton, D., Steele, J. S., & Cisler, J. (2015). The Elicitation and Assessment of Emotional Responding *Sleep and Affect: Assessment Theory and Clinical Implications* (pp. 91-118): Academic Press.

Caruana, N., de Lissa, P., & McArthur, G. (2017). Beliefs about human agency influence the neural processing of gaze during joint attention. *Soc Neurosci, 12*(2), 194-206. doi: 10.1080/17470919.2016.1160953

Caruana, N., & McArthur, G. (2019). The mind minds minds: The effect of intentional stance on the neural encoding of joint attention. *Cognitive, affective & behavioral neuroscience, 19*(6), 1479-1491. doi: 10.3758/s13415-019-00734-y

Choi-Kain, L. W., Finch, E. F., Masland, S. R., Jenkins, J. A., & Unruh, B. T. (2017). What Works in the Treatment of Borderline Personality Disorder. *Curr Behav Neurosci Rep, 4*(1), 21-30. doi: 10.1007/s40473-017-0103-z

Chung, B. Y., Hensel, S., Schmidinger, I., Bekrater-Bodmann, R., & Flor, H. (2020). Dissociation proneness and pain hyposensitivity in current and remitted borderline personality disorder. *Eur J Pain, 24*(7), 1257-1268. doi: 10.1002/ejp.1567

Dammann, G., Rudaz, M., Benecke, C., Riemenschneider, A., Walter, M., Pfaltz, M. C., . . . Gremaud-Heitz, D. J. (2020). Facial Affective Behavior in Borderline Personality Disorder Indicating Two Different Clusters and Their Influence on Inpatient Treatment Outcome: A Preliminary Study. *Front Psychol, 11*, 1658. doi: 10.3389/fpsyg.2020.01658

Denny, B. T., Fan, J., Fels, S., Galitzer, H., Schiller, D., & Koenigsberg, H. W. (2018). Sensitization of the Neural Salience Network to Repeated Emotional Stimuli Following Initial Habituation in Patients With Borderline Personality Disorder. *Am J Psychiatry, 175*(7), 657-664. doi: 10.1176/appi.ajp.2018.17030367

Dixon-Gordon, K. L., Chapman, A. L., Lovasz, N., & Walters, K. (2011). Too upset to think: the interplay of borderline personality features, negative emotions, and social problem solving in the laboratory. *Personal Disord, 2*(4), 243-260. doi: 10.1037/a0021799

Domsalla, M., Koppe, G., Niedtfeld, I., Vollstadt-Klein, S., Schmahl, C., Bohus, M., & Lis, S. (2014). Cerebral processing of social rejection in patients with borderline personality disorder. *Soc Cogn Affect Neurosci, 9*(11), 1789-1797. doi: 10.1093/scan/nst176

Fertuck, E. A., Grinband, J., Mann, J. J., Hirsch, J., Ochsner, K., Pilkonis, P., . . . Stanley, B. (2019). Trustworthiness appraisal deficits in borderline personality disorder are associated with prefrontal cortex, not amygdala, impairment. *Neuroimage Clin, 21,* 101616. doi: 10.1016/j.nicl.2018.101616

Fertuck, E. A., Grinband, J., & Stanley, B. (2013). Facial trust appraisal negatively biased in borderline personality disorder. *Psychiatry Res, 207*(3), 195-202. doi: 10.1016/j.psychres.2013.01.004

Fineberg, S. K., Leavitt, J., Stahl, D. S., Kronemer, S., Landry, C. D., Alexander-Bloch, A., . . . Corlett, P. R. (2018). Differential Valuation and Learning From Social and Nonsocial Cues in Borderline Personality Disorder. *Biol Psychiatry, 84*(11), 838-845. doi: 10.1016/j.biopsych.2018.05.020

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage, 16*(3), 814-821. doi: 10.1006/nimg.2002.1117

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102,* 74-78.

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice Taking from Humans and Machines: An fMRI and Effective Connectivity Study. *Front Hum Neurosci, 10.* doi: ARTN 542

10.3389/fnhum.2016.00542

Gratz, K. L., Richmond, J. R., Dixon-Gordon, K. L., Chapman, A. L., & Tull, M. T. (2019). Multimodal assessment of emotional reactivity and regulation in response to social rejection among self-harming adults with and without borderline personality disorder. *Personal Disord, 10*(5), 395-405. doi: 10.1037/per0000334

Gunderson, J. G., Fruzzetti, A., Unruh, B., & Choi-Kain, L. (2018). Competing Theories of Borderline Personality Disorder. *J Pers Disord, 32*(2), 148-167. doi: 10.1521/pedi.2018.32.2.148

Gunderson, J. G., Herpertz, S. C., Skodol, A. E., Torgersen, S., & Zanarini, M. C. (2018). Borderline personality disorder. *Nat Rev Dis Primers, 4,* 18029. doi: 10.1038/nrdp.2018.29

Gunderson, J. G., Stout, R. L., McGlashan, T. H., Shea, M. T., Morey, L. C., Grilo, C. M., . . . Skodol, A. E. (2011). Ten-year course of borderline personality disorder: psychopathology and function from the Collaborative Longitudinal Personality Disorders study. *Arch Gen Psychiatry, 68*(8), 827-837. doi: 10.1001/archgenpsychiatry.2011.37

Henco, L., Diaconescu, A. O., Lahnakoski, J. M., Brandi, M. L., Hormann, S., Hennings, J., . . . Mathys, C. (2020). Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS Comput Biol, 16*(9), e1008162. doi: 10.1371/journal.pcbi.1008162

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., & Montague, P. R. (2008). The rupture and repair of cooperation in borderline personality disorder. *Science, 321*(5890), 806-810. doi: 10.1126/science.1156902

Kleindienst, N., Jungkunz, M., & Bohus, M. (2020). A proposed severity classification of borderline symptoms using the borderline symptom list (BSL-23). *Borderline Personal Disord Emot Dysregul, 7,* 11. doi: 10.1186/s40479-020-00126-6

Koenigsberg, H. W., Denny, B. T., Fan, J., Liu, X., Guerreri, S., Mayson, S. J., . . . Siever, L. J. (2014). The neural correlates of anomalous habituation to negative emotional pictures in borderline and avoidant personality disorder patients. *Am J Psychiatry, 171*(1), 82-90. doi: 10.1176/appi.ajp.2013.13070852

Krauch, M., Ueltzhoffer, K., Brunner, R., Kaess, M., Hensel, S., Herpertz, S. C., & Bertsch, K. (2018). Heightened Salience of Anger and Aggression in Female Adolescents With Borderline Personality Disorder-A Script-Based fMRI Study. *Front Behav Neurosci, 12,* 57. doi: 10.3389/fnbeh.2018.00057

Kraus, A., Valerius, G., Seifritz, E., Ruf, M., Bremner, J. D., Bohus, M., & Schmahl, C. (2010). Script-driven imagery of self-injurious behavior in patients with borderline personality disorder: a pilot FMRI study. *Acta Psychiatr Scand, 121*(1), 41-51. doi: 10.1111/j.1600-0447.2009.01417.x

Li, C. S., Kosten, T. R., & Sinha, R. (2005). Sex differences in brain activation during stress imagery in abstinent cocaine users: a functional magnetic resonance imaging study. *Biol Psychiatry, 57*(5), 487-494. doi: 10.1016/j.biopsych.2004.11.048

Li, C. S., & Sinha, R. (2006). Alexithymia and stress-induced brain activation in cocaine-dependent men and women. *J Psychiatry Neurosci, 31*(2), 115-121.

Liebke, L., Koppe, G., Bungert, M., Thome, J., Hauschild, S., Defiebre, N., . . . Lis, S. (2018). Difficulties with being socially accepted: An experimental study in borderline personality disorder. *J Abnorm Psychol, 127*(7), 670-682. doi: 10.1037/abn0000373

Limberg, A., Barnow, S., Freyberger, H. J., & Hamm, A. O. (2011). Emotional vulnerability in borderline personality disorder is cue specific and modulated by traumatization. *Biol Psychiatry, 69*(6), 574-582. doi: 10.1016/j.biopsych.2010.10.024

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proc Natl Acad Sci U S A, 98*(20), 11832-11835. doi: DOI 10.1073/pnas.211415698

Miano, A., Fertuck, E. A., Arntz, A., & Stanley, B. (2013). Rejection sensitivity is a mediator between borderline personality disorder features and facial trust appraisal. *J Pers Disord, 27*(4), 442-456. doi: 10.1521/pedi_2013_27_096

Pfeiffer, U. J., Schilbach, L., Timmermans, B., Kuzmanovic, B., Georgescu, A. L., Bente, G., & Vogeley, K. (2014). Why we interact: on the functional role of the striatum in the subjective experience of social interaction. *Neuroimage, 101,* 124-137. doi: 10.1016/j.neuroimage.2014.06.061

Polgar, P., Fogd, D., Unoka, Z., Siraly, E., & Csukly, G. (2014). Altered social decision making in borderline personality disorder: an Ultimatum Game study. *J Pers Disord, 28*(6), 841-852. doi: 10.1521/pedi_2014_28_142

Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prevot, L., & Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B-Biological Sciences, 374*(1771). doi: ARTN 20180033
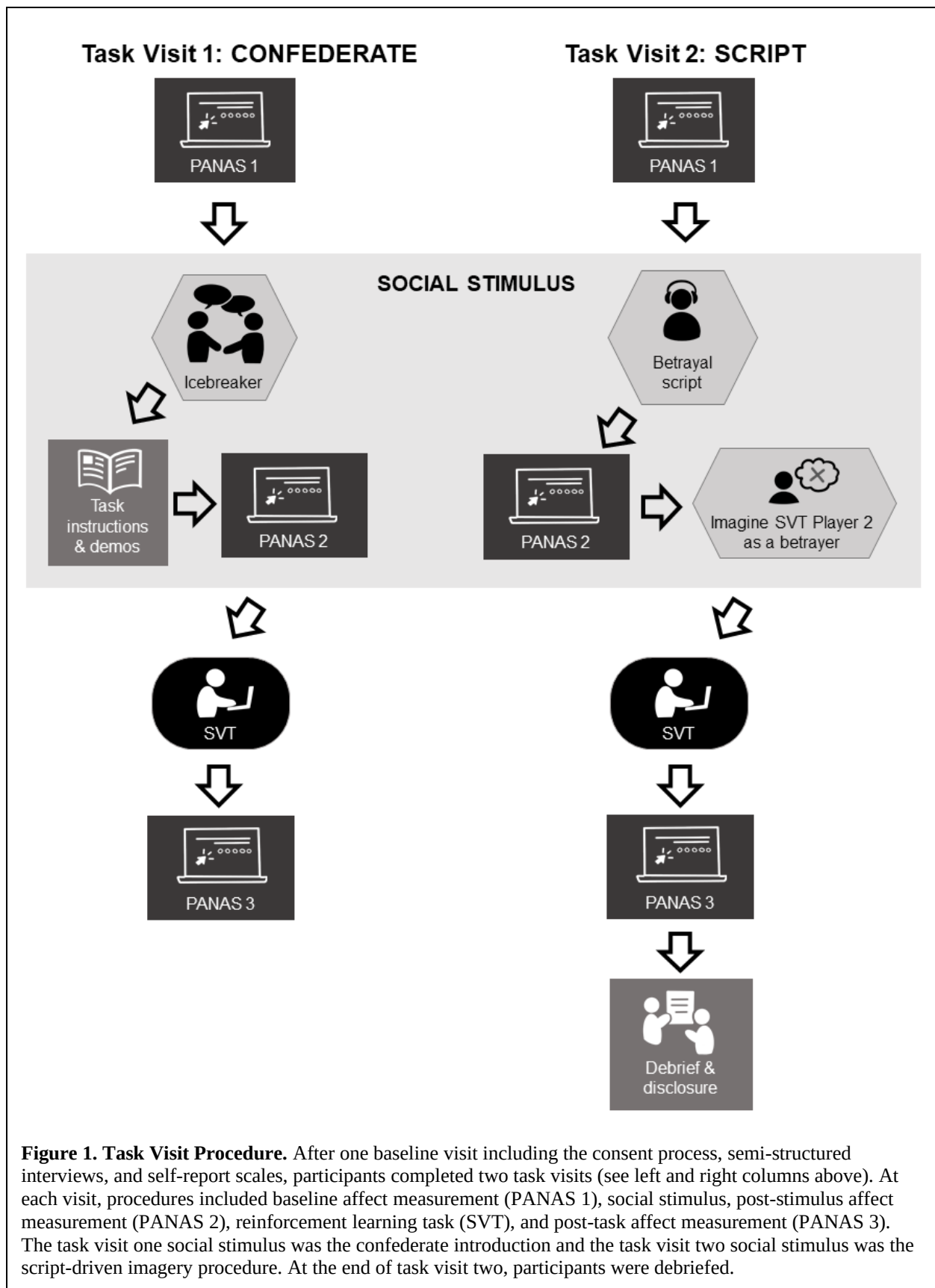
10.1098/rstb.2018.0033

Rice, K., & Redcay, E. (2016). Interaction matters: A perceived social partner alters the neural processing of human speech. *Neuroimage, 129*, 480-488. doi: 10.1016/j.neuroimage.2015.11.041

Robins, C. J. (1988). Development of experimental mood induction procedures for testing personality-event interaction models of depression. *Journal of Clinical Psychology, 44*(6), 958-963. doi: 10.1002/1097-4679(198811)44:6<958::aid-jclp2270440618>3.0.co;2-d

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences, 36*(4), 393-414. doi: 10.1017/S0140525X12000660

Schindler, S., & Kissler, J. (2016). People matter: Perceived sender identity modulates cerebral processing of socio-emotional language feedback. *Neuroimage, 134*, 160-169. doi: 10.1016/j.neuroimage.2016.03.052

Schulze, L., Schmahl, C., & Niedtfeld, I. (2016). Neural Correlates of Disturbed Emotion Processing in Borderline Personality Disorder: A Multimodal Meta-Analysis. *Biol Psychiatry, 79*(2), 97-106. doi: 10.1016/j.biopsych.2015.03.027

Sevgi, M., Diaconescu, A. O., Tittgemeyer, M., & Schilbach, L. (2015). Social Bayes: Using Bayesian Modeling to Study Autistic Trait-Related Differences in Social Cognition. *Biol Psychiatry*. doi: 10.1016/j.biopsych.2015.11.025

Sinha, R., & Tuit, K. (2012). *Imagery Script Development Procedures Manual*.

Smarr, K. L., & Keefer, A. L. (2011). Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). *Arthritis Care Res (Hoboken), 63 Suppl 11*, S454-466. doi: 10.1002/acr.20556

Team, R. C. (2014). R: A Language and Environment for Statistical

Computing. Vienna, Austria.

Ueltzhoffer, K., Herpertz, S. C., Krauch, M., Schmahl, C., & Bertsch, K. (2019). Whole-brain functional connectivity during script-driven aggression in borderline personality disorder. *Prog Neuropsychopharmacol Biol Psychiatry, 93*, 46-54. doi: 10.1016/j.pnpbp.2019.03.004

Unoka, Z., Fogd, D., Fuzy, M., & Csukly, G. (2011). Misreading the facial signs: specific impairments and error patterns in recognition of facial emotions with negative valence in borderline personality disorder. *Psychiatry Res, 189*(3), 419-425. doi: 10.1016/j.psychres.2011.02.010

Unoka, Z., Seres, I., Aspan, N., Bodi, N., & Keri, S. (2009). Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. *J Pers Disord, 23*(4), 399-409. doi: 10.1521/pedi.2009.23.4.399

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol, 54*(6), 1063-1070. doi: 10.1037//0022-3514.54.6.1063

Zanarini, M., Frankenburg, F. R., Chauncey, D. L., & Gunderson, J. G. (1987). The diagnostic interview for personality disorders: Interrater and test-retest reliability. *Compr Psychiatry, 28*(6), 467-580.

Zhang, M., Dumas, G., Kelso, J. A., & Tognoli, E. (2016). Enhanced emotional responses during social coordination with a virtual partner. *Int J Psychophysiol, 104*, 33-43. doi: 10.1016/j.ijpsycho.2016.04.001

TABLE 1. Demographics

	BPD (*n* = 21)	HC (*n* = 20)
Age in years M ± SD	33.2 ± 10.4	38.4 ± 13.3
Gender		
Female	16 (76%)	15 (75%)
Non-binary or genderqueer	2 (10%)	1 (5%)
Male	3 (14%)	4 (20%)
Race		
Asian	3 (14%)	4 (20%)

TABLE 2. Self-report scores

	BPD (*n* = 21)	HC (*n* = 19)	*U/t*	*p*	Cohen's *d*
BSL-23	1.58 (0.75)	0.21 (0.15)	*U* = 380	p < 0.001	*d* = 2.61
BDI	28.1 (11.7)	4.2 (5.6)	*U* = 387.5	p < 0.001	*d* = 2.60

Notes: Results shown as mean (SD); BSL-23 = Borderline Symptom Scale; BDI = Beck Depression Inventory II

Figure 1. Task Visit Procedure. After one baseline visit including the consent process, semi-structured interviews, and self-report scales, participants completed two task visits (see left and right columns above). At each visit, procedures included baseline affect measurement (PANAS 1), social stimulus, post-stimulus affect measurement (PANAS 2), reinforcement learning task (SVT), and post-task affect measurement (PANAS 3). The task visit one social stimulus was the confederate introduction and the task visit two social stimulus was the script-driven imagery procedure. At the end of task visit two, participants were debriefed.
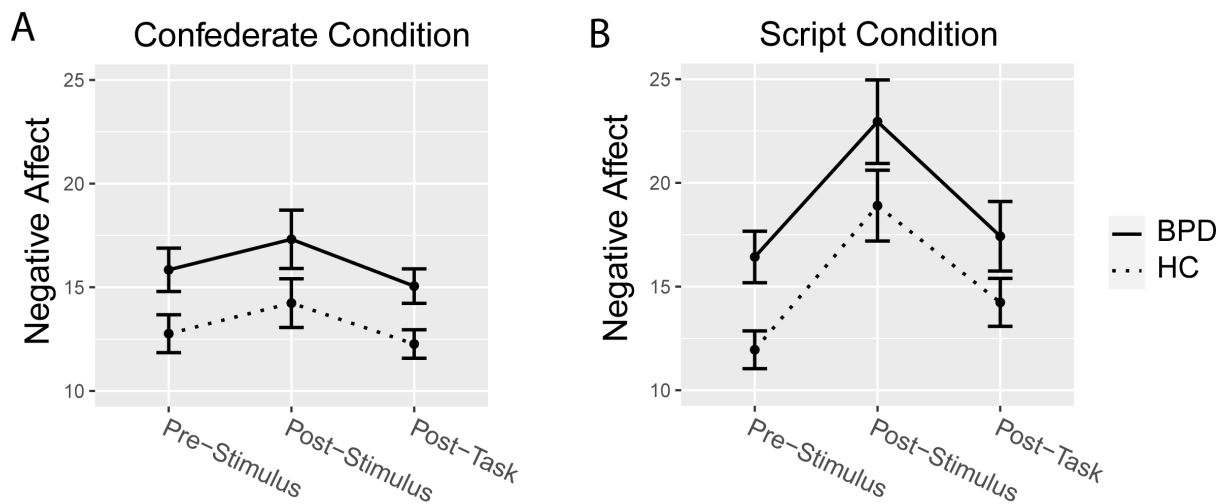
Figure 2: Differences in Negative Affect by Condition. PANAS scores in the confederate condition (A) and the script condition (B) reveal that the scripts had a greater impact on negative affect than meeting the confederate.

Notes: PANAS negative affect scores range from 10-50. Bars indicate standard error.
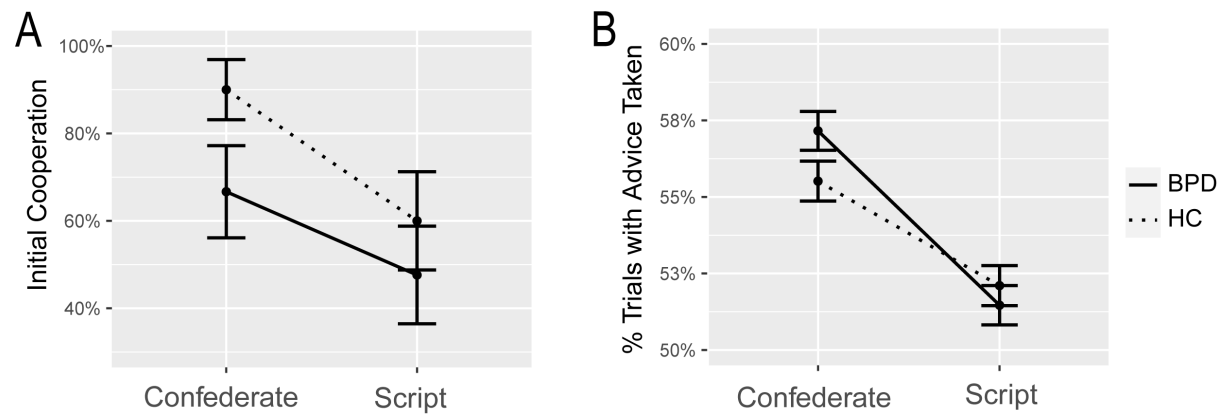
Figure 3. Advice-Taking. In the script condition, both groups accepted advice less than in the confederate condition: see initial cooperation (**3A**) and % advice-taking (**3B**). For % advice-taking, the effect was stronger in BPD than HC. Error bars represent standard error of the mean.

TABLE 3. Trust behavior regression results

	Estimate	Std. Error	z	p
*Took advice ~ Recent Betrayal * Group * Condition*				
Recent Betrayal	-0.536	0.052	-10.294	< 0.001
Condition (Script)	-0.151	0.050	-3.014	0.002
Group (BPD) * Betrayal	-0.168	0.074	-2.263	0.023
Condition (Script) * Betrayal	-0.175	0.073	-2.380	0.017

TABLE 4. Individual differences in trust behavior

	Estimate	Std. Error	z	p
*Took advice ~ BSL-23 * Condition*				

Supplemental Materials

Appendix A: Betrayal Script

It's a gloomy midsummer afternoon. You're sitting in your apartment, waiting for your ex to come pick their stuff up and move out. Your stomach is in a knot. You hear the door open and you see your ex. Your heart beats faster. Then you see your best friend trailing behind them. You are shocked. Your mind starts racing. "What're you doing here?", you ask. They respond with silence. The muscles in your face begin to tighten. You can't believe it: your best friend is siding with your abusive ex. Are you fucking kidding me. Your hands start trembling. They start removing furniture. "How dare you!", you scream. Now your whole body shakes. "You're supposed to be MY friend!" They continue to ignore you. You feel you have no choice but to sit there and watch your whole life fall apart in front of you. You gasp for air. You feel like you are exploding from the inside out. You want to scream or to strike someone. You want to hurt your so-called best friend.