A move-by-move paradigm for the computational characterization of attachment style and personality disorder

Federico Mancinelli[a,*], Tobias Nolte[b,c], Julia Griem[b,d], London Personality and Mood Disorder Research Consortium, Terry Lohrenz[e], Janet Feigenbaum[b], Brooks King-Casas[e], P. Read Montague[e], Peter Fonagy[c], and Christoph Mathys[a,f]

[a]*Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy*
[b]*Research Department of Clinical, Educational and Health Psychology, University College London, London, UK*
[c]*Anna Freud National Centre for Children and Families, London, UK*
[d]*Division of Psychology and Language Sciences, University College London, UK*
[e]*Fralin Biomedical Research Institute, Virginia Polytechnic Institute and State University, USA*
[f]*Interacting Minds Centre, Aarhus University, Aarhus, Denmark*

Abstract

A current direction of personality disorder research strives to identify key behavioural, cognitive, and ultimately computational facets of patient functioning via the use of engaging social paradigms. Thus far, few such paradigms have been put forward. Here, we introduce a novel task in which subjects interact with previously unknown virtual partners in a turn-taking paradigm akin to a dance, and subsequently report on their experience with each. The partners' "personalities" differ in the nature and extent of their reactions to the inter-personal distance kept by participants. We show that the plurality of measures produced may help further characterize attachment style and borderline personality disorder (BPD) symptoms. Higher scores on our measures of attachment anxiety, avoidance, and BPD symptoms were all linked to a general negative appraisal of all the interpersonal experiences. Further, the personalities of the partners encountered mattered: for instance, negative appraisal of a partner who displayed the most biasedly negative range of moods was tied with attachment anxiety and BPD symptoms. Finally, our analyses of proxemics data underscored slower movement initiation from anxiously attached individuals throughout all virtual interactions, whereas BPD symptoms were tied with a tendency to react with further distancing from a partner which is too close.

1. Introduction

Borderline personality disorder (BPD) is a highly prevalent and debilitating clinical condition, which spans a diverse range of symptoms in the interpersonal, affective, cognitive, and behavioural domains (circa 0.5 - 2.5 % of the population are affected; see, e.g. Maier et al., 1992; Gunderson et al., 2011; Gunderson, 2009; Skodol et al., 2002; Leichsenring et al., 2011). Symptoms of BPD have been shown to correlate very highly ($r \sim 0.8$) with the general psychopathology factor in the bi-factor analysis of two major epidemiology trials (Gluschkoff et al., 2021). However, our understanding of the cognitive and emotional mechanisms underlying BPD is still limited, and diagnosis relies almost exclusively on clinical interviews and questionnaires because quantifiable cognitive and biological markers of the disorder are lacking. In the present study, we address these shortcomings by introducing a novel task designed to allow for the 'cognitive-emotional fingerprinting' of subjects' reactions to their interactions with (virtual) others.

A decisive consideration in the design of the task and the associated data analyses was that attachment disturbances and the ensuing continual patterns of interpersonal dysfunction are defining themes of the aetiology and maintenance of BPD, respectively (see e.g. Agrawal et al., 2004; Gunderson, 2007; Fossati et al.,

*Corresponding author (federico.mancinelli@sissa.it)

1999; Johansen et al., 2004; Lieb et al., 2004). Patients frequently have intense and unstable intimate relationships and typically exhibit shifts between idealisation and devaluation of the other as well as aggression and extreme distress at perceived threats of abandonment (APA, 2013; WHO, 2004).

Research in experimental psychology pays due attention to the identification of core cognitive endo- and eco-phenotypes of BPD underpinning the interpersonal domain, for instance in so-called "static" tasks such as mental state discrimination (e.g. Fertuck et al., 2009; Frick et al., 2012; Anupama et al., 2018; Berenson et al., 2018), facial emotion recognition (e.g. Lowyck et al., 2016; Ritzl et al., 2018), or reactivity to emotion induction (Renneberg et al., 2005). There have also been growing efforts to investigate aspects of aberrant appraisal of social interactions, such as in paradigms of social exclusion and rejection (Domsalla et al., 2014; De Panfilis et al., 2015), idealisation and devaluation (Michael et al., 2021), or behavioural trust games (King-Casas et al., 2008; Unoka et al., 2009; Xiang et al., 2012). Other accounts have provided interesting characterisations of differentiable biases in social vs. non-social (i.e. physical) learning in BPD patients (Behrens et al., 2008; Fineberg et al., 2018b; Henco et al., 2020).

However, there is a growing agreement that novel paradigms are needed to elicit a more proximal range of dysfunctional appraisal mechanisms (and ensuing behaviours) that arise ecologically in BPD - moving towards settings in which participants engage with the closer "you" rather than the distant "her/him" (Schilbach et al., 2013; Fineberg et al., 2017). A promising approach, which we adopt here, is to use virtual environments in which social interactions take place with avatars (McCall, 2015; Fineberg et al., 2017; Michael et al., 2021; Sevgi et al., 2020). We use the term avatar to refer to computer-generated agents whose behaviour is designed to be human-like but who are not necessarily human-like in appearance. In this setting, the experimenter is free to design the participant's behavioural repertoire so that it becomes measurable; define the social algorithms of the interacting virtual entities; and shape the situation or experimental manipulation so that it may effectively probe the internal computing models - including cognitive or affective biases - which cause and perpetuate interpersonal dysfunction (see Barakova et al., 2009, for an application to autism). When adequate levels of engagement are achieved, virtual environments constitute a powerful device to elicit the behaviours (Bailenson et al., 2003) and engage neural circuitry (Mathiak & Weber, 2006) which would normally be in place when experiencing similar situations in real life.

As a step in this direction, we developed the "dancing task", a paradigm that enabled us to describe and decompose the appraisal of virtual interpersonal interactions. This allowed us to relate both the reported quality of this experience and objective measures of behaviour to the spectrum of borderline symptomatology and attachment styles. The dancing task was inspired by a seminal but somewhat underexploited approach which uses visual animations to elicit attributions of actions, interactions and mental states to others (Heider & Simmel, 1944). It makes use of minimal avatars (smiley-frowny faces) in a 2D space (a computer screen), where subjects get to know (or rather "dance" with - by controlling their own avatar) four different partners, which vary in their "personalities" (i.e., in the bias and range of their emotion expression which lies on a continuum from smiling to frowning). Subjects were asked to remember their experience as clearly as possible. The adequacy of this choice of minimal smiley-frowny avatar is supported by evidence that similar patterns of neural activity related to emotional processing occur when people are exposed even just to simple text emoticons (Yuasa et al., 2006; Aldunate & González-Ibáñez, 2017). Subjects were told that their partners' emotions would vary according to how the dance unfolds. Specifically, they were informed that in order to keep their partner happy (i.e., smiling), they would need to keep a "correct" (i.e., comfortable to the partner) interpersonal distance to it as they rhythmically moved across the screen in a back-and-forth manner. This dependence of the avatar's mood on physical distance was motivated by the known difficulties of BPD patients with keeping the right psychological distance from others and by several studies showing that BPD is associated with altered responses to emotional stimuli (with mixed results – see e.g., Renneberg et al., 2005; Matzke et al., 2014; Bertsch et al., 2018; Mitchell et al., 2014). However, the dancing task goes beyond previous efforts to elicit emotional responses to emotional stimuli. It represents a step forward in terms of ecological validity because there is a clear difference between a situation where a subject passively appraises a facial expression and one in which they have a role in causing it. Furthermore, in previous tasks subjects were not induced to form a mental representation of another entity by repeated, emotionally charged interaction.

A first part of our analyses was focused on the experience retained from the dances (assessed via questionnaire). We then linked this experience to the spectrum of borderline symptoms and attachment styles. We investigated both passive aspects of the experience, such as the extent to which each dancer was liked, or could be trusted, after the interaction, as well as active and action-dependent aspects, such as how much was the dancer made happy or sad during the dance. A second set of analyses focused on proxemic variables, such as move-by-move click distances and reaction times throughout the dances.

Our primary hypothesis concerned an increasingly negatively biased appraisal of the overall dancing experience as BPD symptoms severity and attachment difficulties increased. We sought to add granularity to this by identifying facets of the experience which are particularly predictive of symptom severity when *specific* personalities (i.e., dancers) are encountered. Most BPD symptoms emerge in interpersonal contexts. Therefore, by systematically characterizing the reported experience from interactions with varying dancer personalities, our aim was to develop a task and associated measures that amount to a kind of cognitive-emotional fingerprint of each subject. Ideally, such a fingerprint should be diagnostic and possibly even prognostic at the mechanistic level, which would take us beyond symptom-based classification and might also give new scope for treatment.

2. Methods

2.1. Participants and procedure

Fourty-eight subjects diagnosed with borderline personality disorder and 38 healthy controls took part in the study. Participants diagnosed with BPD were recruited from specialist personality disorder services across various London mental health trusts. The diagnosis of BPD was confirmed using the Structured Clinical Interview for DSM-IV (SCID-II; First et al., 1997). Individuals with a history of psychotic episodes, severe learning disability or neurological illness/trauma were excluded. Healthy control participants were recruited from the community. They did not have a history of mental illness or neurological illness/trauma and did not have any current diagnosis. The absence of personality disorder in healthy controls was confirmed by screening participants with the Standardized Assessment of Personality, Abbreviated Scale (SAPAS; Moran et al., 2003). Any individual scoring above 4 on the SAPAS was subsequently interviewed with the SCID-II and excluded if they scored above threshold on any personality disorder. All participants were included on the basis of English language fluency. Participants attended research appointments at University College London. All participants provided signed informed consent. The study was approved by the Research Ethics Committee for Wales (REC reference number 12/WA/0283). One control subject and three patients were excluded from the analyses. Two patients and one control participant were excluded due to missing questionnaire data. One more patient was excluded on account of spending an outlying amount of time with just one partner (more than 75% of the total task time). For two more control subjects, the ECR-R scales were not available. These missing data left us with 82 subjects (37 controls and 46 patients) for the analyses involving PAI-BOR measures, and 80 subjects (35 controls and 46 patients) for the analyses involving ECR-R scales.

2.2. Questionnaires

2.2.1. Personality Assessment Inventory for Borderline Traits (PAI-BOR)

The PAI-BOR is a self-report questionnaire assessing traits associated with BPD (Morey, 1991). Across 24 items, participants are asked to indicate how much each question describes them from 0 ("False") to 3 ("Very True"). Combining all items gives a total score (PAI-BOR). Additionally, there are four subscales relating to core BPD features: affective instability (PAI-BOR-A), negative relationships (PAI-BOR-N), identity problems (PAI-BOR-I), and non-suicidal self-harm (PAI-BOR-S). PAI-BOR-S merges impulsive behaviours and self-harm. For all scales, a higher score indicates more severe pathology.

2.2.2. Experiences of Close Relationships-Revised (ECR-R)

The ECR-R is a self-report questionnaire measuring adult attachment tendencies towards romantic partners in terms of how anxious or avoidant they are (Fraley et al., 2000). Subjects answered 36 questions asking them to indicate how much they agree with a given item on a range from 1 ("Strongly Disagree") to 7 ("Strongly Agree"). This results in scores for two subscales: Anxious-Attachment and Avoidant-Attachment. A higher score represents a higher level of anxious or avoidant attachment.

2.3. Task design

The dancing task consisted of a JavaScript-coded game (available online at `https://ba5r373hms.cognition.run/`). The game involved a series of dancing episodes between the subject's avatar and each of four virtual partners, all shown as circular smiley-frowny faces on a blank canvas (the 'dance floor'). The four partners differed in their personalities, defined by the individual range of moods they were able to express through their mouth and eyes. Partners could be identified by their colours. The subject's avatar's facial expression (i.e., the expression of the smiley face representing the participant's position in the virtual space) was kept neutral.

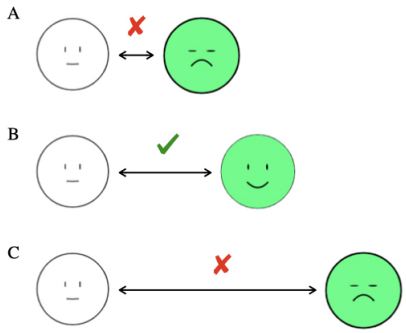

Figure 1: **The evolution of mood as a function of distance.** This figure provides a simple illustration of the relationship between dancing distance and the evolution of mood. The transition between facial expressions is governed by a differential equation, which makes it gradual (see Supplementary Materials). In **A** the subject is too close to the partner; in **B**, the subject strikes a good distance from the partner, corresponding to a certain interval (i.e., between 55 and 200 pixels); in **C**, the subject is too far.

2.3.1. Task structure

After registering their preferred username, subjects faced a short training dance (1 min.) to familiarise themselves with moving their avatar (which we call S for brevity). Once this was completed, four coloured circles (the new dancing partners) made their appearance. When these were not selected, they were simply shown as plain, numbered (1-4) circles, each of a different colour. Subjects simply pressed the corresponding key (1-4) to select a partner for a dance. When a partner was selected (we call the selected partner P) it turned from a plain circle into a smiley-frowny face. Subjects could then see the partner's facial expression (neutral before the first move) and the dance began when the subject first moved their avatar. From this point onwards, P's mood was a function of distance to the subject's avatar. Specifically, the update of the selected partner's mood during the dance was determined by a simple differential equation (described in Figure 1 in SMs), which dictated that mood improved when S struck a *good distance* from P (not too close, nor too far), and deteriorated otherwise (Figure 1). Dances could be interrupted at any moment (after a minimum of $3s$) by pressing the space bar. Subjects knew that they must dance with all partners at least once; to enforce this, partners could not be re-selected before all had been given one dance first. However, once all partners had been given one dance, subjects were free to re-select whichever dancer they preferred. Subjects had 14 minutes to get to know all partners, after which they filled in a questionnaire about their impressions of each.

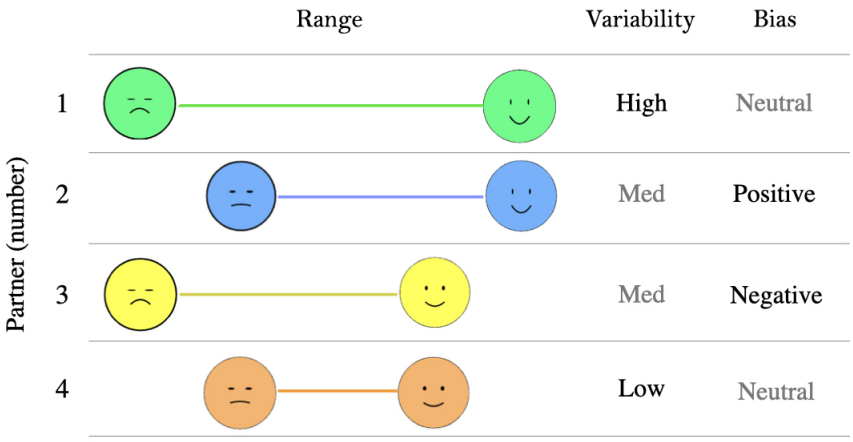

Figure 2: **Personalities of partners.** Dancing partners differ in the breadth and bias of their facial expressions (columns "Variability" and "Bias" respectively). Partner 1 (top row) draws their facial expression from the full range - by contrast, partner 4's range of expressiveness is symmetrically reduced. Both partners 1 and 4 have "Neutral" biases in the sense that their expressive range is symmetrical. Conversely, partners 2 and 3 have somewhat reduced ranges of expressiveness, but are biased. Partner 2 can reach the full extreme on the positive affect side, but not on the negative side, and vice-versa for partner 3.

2.3.2. Dancing task questionnaire

We developed a questionnaire (Dancing Task Questionnaire; DTQ) which was filled in just after the dances were completed. The questionnaire consisted of the same set of nine items concerning each of the four partners (36 items in total). Subjects responded via a JavaScript visual analogue scale which allowed for finely graded responses (coded as numbers from 0 to 100). The items are listed below. Next to each item (in quotes, italic) we indicate the label by which we refer to it in what follows.

1. How much did you like this dancer? ("*Likeable*")
2. How much did you trust this dancer? ("*Trustworthy*")
3. Do you think you made this dancer happy? ("*Made Happy*")
4. Do you think you made this dancer sad? ("*Made Sad*")
5. How much did you get irritated or annoyed with this dancer? ("*Irritating*")
6. How much effort did you invest in understanding which distance this dancer liked? ("*Effort*")
7. How unpredictable was their dance? ("*Unpredictable*")
8. Did you feel that their mood was unstable? ("*Unstable*")
9. Did you feel this dancer's mood depended on what you were doing? ("*Locus*")

An earlier version of the questionnaire only contained 8 questions per partner, so for some subjects (27), "*Trustworthy*" ratings (item 2) were unavailable. We imputed the missing values via "3 nearest neighbours", a good trade-off between accuracy and preservation of data structure (Beretta & Santaniello, 2016).

2.4. Analyses

We initially carried out an exploratory analysis of our clinical questionnaires. This was done obtaining Chronbach's α measures to tap the internal consistency of each scale, and subsequently performing sparse canonical correlations analysis (sCCA; Witten et al., 2009, PMA package) across all clinical subscales, to explore the relations existing among them within our sample. The bulk of our analyses concerned the ways in which the impression gained of each partner, measured via the dancing task questionnaire, varied with BPD symptoms severity (PAI-BOR) and attachment style (ECR-R). The DTQ was designed so that the same set of 9 questions was asked about each dancing partner. Thus, response patterns could emerge (1) regardless of the partner a particular question is asked about, or (2) only when targeting a particular subset of partners, revealing an influence of their personality. We therefore first analyzed questionnaire responses

marginalizing across partners, and studied the covariance structure of these partner-independent measures, identifying dimensions of principal variations through Principal Components Analysis (PCA; Pearson, 1901). The rationale of this PCA-based analysis was that scores along these partner-independent dimensions will reflect prior notions about people *in general* (regardless of personality) and, as a DTQ pre-processing step, they would inform us as to the underlying dimensionality of the DTQ responses. We used parallel analysis (custom code) to establish the number of meaningful underlying components (Horn, 1965). Once this step was completed, we proceeded by measuring the effect of partner personality, by projecting the scores assigned to each partner along the principal dimensions found, and computing Bonferroni-Holmes corrected pairwise comparisons. In a third and final step, we explored the responses to individual items and their relationship with PAI-BOR and ECR-R sub-scales. Here, we adopted a model-based approach using sparse (i.e., regularized) canonical correlations analysis (sCCA; Witten et al., 2009, PMA package) to determine the main mode through which DTQ and clinical questionnaire measures (PAI-BOR and ECR-R) relate to each other. This approach differs from the former in that it actively looks for a critical linear combination of (1) questionnaire measures and (2) all ratings for items in the DTQ, to maximise their correlation, thus providing a measure for the relative contribution of each clinical subscale and DTQ item towards maximising the relationships between the two sets. sCCA has the advantage that it is straightforward to interpret and provides a principled and robust way to quantify links between dancing items and questionnaire scores. In other words, we used regularized CCA for robustness and to avoid overfitting. For completeness, in the Supplementary Material, we also report pairwise correlations with Bonferroni-Holmes corrected p-values.

2.4.1. Outcome measures and hypotheses

Our primary hypothesis was that the experience retained from dances would be increasingly negative with higher PAI-BOR symptom severity or the degree of insecure attachment (i.e. both features of anxiety and avoidance). While we presumed that partners 1 and 3 might be particularly relevant on account of the more variable and biased negative range of moods respectively, we had no strong hypotheses about which particular item(s) would be predictive of which symptoms. This hypothesis was reflected in our modelling approach (sCCA). Our quantitative analyses of overt behaviour and proxemics (i.e. the microscopic decisions about how and where to go as a response to a partner's move) included (1) putative proxies for indecision (i.e. average reaction times; Laming, 1968), (2) proxies of preference for interpersonal distance and reaction to space intrusion (similar to, e.g. Bailenson et al., 2003) and (3) the proportion of time spent with each partner throughout the task. We considered the time spent during the exploratory phase (in which subjects must dance with all partners at least once) and the subsequent phase, in which choice of the next dancing partner is unconstrained. Rather than consider measures (1) and (2) in relation to each of the four partners independently, we first condensed them using dimensionality reduction (PCA) as in the first DTQ analysis, a pre-processing step which we used to ensure that there is meaningful variability of these measures across partners.

2.4.2. sCCA predictive performance

We assessed the out-of-sample performance of the sCCA model (i.e., the expected strength of the association between dancing questionnaire and PAI and ECR-R scales) via cross-validation. We split our dataset into five folds. For each of five iterations, one fold was held out as the remaining four were merged and used as training set. Here, we used the native permutation scheme implemented in the PMA package to extract the best penalization parameters (denoted λ; i.e., the L1 norm upper-bounds on the CCA weight vectors), which were then used to compute a correlation coefficient between latent dimensions in the held-out set. Out-of-sample correlations were averaged to yield stability. Lastly, we took the median of five repetitions.

3. Results

3.1. Clinical questionnaires

We found that both the PAI-BOR and ECR-R questionnaire sub-scales have very good internal consistencies (Cronbach's α; PAI-BOR-A: 0.92, PAI-BOR-N: 0.82, PAI-BOR-I: 0.83, PAI-BOR-S: 0.90; ECR-R-Anxiety: 0.93, ECR-R-Anxiety: 0.93). We then performed sCCA on the full dataset, as anticipated in

PAI	sCCA weights
PAI-BOR-A (Affective Instability)	0.46
PAI-BOR-N (Negative relationships)	0.49
PAI-BOR-I (Identity problems)	0.66
PAI-BOR-S (Self-harm)	0.32
ECR-R	
Anxiety	0.98
Avoidance	0.17

Table 1: **Relationship between attachment scores and PAI borderline scales**. This table quantifies the relationship found through sCCA between PAI sub-scales (top four rows), and attachment style as measured by ECR-R along dimensions of anxiety and avoidance (bottom two rows). The relationship found was strong (R = 0.84), with only minor penalisations on both sides (PAI: $\lambda = 0.97$, ECR-R: $\lambda = 0.79$), and highlighted a prominent role of attachment anxiety in relation to borderline features - with substantial focus on identity problems (PAI-BOR-I).

methods, to link the primary latent dimensions of (1) the four PAI-BOR subscales and (2) the two ECR-R sub-scales. The foreshadowed relationship between the two latent dimensions was found to be very strong (R = 0.84). This analysis highlights a link between attachment anxiety (ECR-R-Anxiety) and identity problems (PAI-BOR-I) in our sample, with lower contributions from the other PAI-BOR subscales, and a substantially lower contribution from attachment avoidance. Results from this analysis are illustrated in table 1

3.2. Partner-independent analyses

We first identified the principal dimensions on marginalised items (i.e., items obtained by summing ratings across partners, thereby eliminating the effect of partner personality), to then test our primary hypothesis of PAI-BOR and ECR-R scales as predictors of a general negative appraisal of the interaction with all partners. Our parallel analysis identified three meaningful components of our PCA on marginalized DTQ items. These components cumulatively explained a variance of 69.9% (1^{st}component : 36%, 2^{nd}component : 18.5%, 3^{rd}component : 15.5%). The resulting principal dimensions were readily interpretable (see table 3, or figure 3 for a depiction of the first two components). In the first dimension, positive scores were associated with positively oriented items, i.e. *"Likeable"* (contributing 16.5% for this dimension; we hereafter report the relative contribution, in parentheses, next to each item), *"Trustworthy"* (16.4%), *"Made Happy"* (13.6%) whereas negative scores were associated with negatively oriented items, i.e. *"Made Sad"* (8.6%), *"Irritating"* (19%), *"Unpredictable"*(8.7%), *"Unstable"*(16.0%). The second dimension saw positive contributions from all items - and despite covarying most strongly with items *"Effort"* (31.7%) and *"Unpredictable"*(24.3%), it can be conservatively interpreted as reflecting a general tendency to give high or low answers in the DTQ. The third dimension covaried with the subjective tendency to report making partners sad (and not happy) and feeling responsible for it - as the most contributing items were *"Made Sad"*(29%) and *"Locus"*(37%). Thus, the primary dimension embodies a general or summary appraisal of each partner (i.e., a positive or negative impression) while the second dimension conveys a tendency to give overall high or low ratings, in which however most prominent are the perception of effort exerted into making the partner happy (item 6) and of how unpredictable they had been experienced to be (item 7). Finally, the third dimension, appears to reveal a retainment to have caused partners to be sad/not happy. The first dimension (which we herafter refer to as "summary appraisal") anti-correlated with summed PAI-BOR scores and all subscales (PAI-BOR: $r = -0.37$, $p < 0.001$, $CI_{95\%}[-0.55, -0.17]$, subscales: all $r < -0.28$, all $p_{adj} < 0.01$), and both ECR-R subscales, though the strength of the association was noteworthier for the anxiety sub-scale (anxiety: $r = -0.46$, $p < 0.001$, $CI_{95\%}[-0.61, -0.26]$; avoidance: $r = -0.33$, $p = 0.003$, $CI_{95\%}[-0.51, -0.11]$). There were no meaningful correlations involving the second or third PCA dimensions (all corrected p-values ≥ 0.1). In sum, higher PAI-BOR and ECR-R scores determine a more negative experience for all dancing interactions, with anxiety scores playing a prominent role, but are not appreciably related to the magnitude of scores assigned (dim.2), or retaining of being responsible to have caused partners to be sad/not happy (dim.3).

Partner	PAI-BOR	ECR-R	
		Anxiety	Avoidance
1	-0.20	-0.21	-0.15
2	-0.14	-0.15	-0.23
3	-0.32^{*}	-0.33^{*}	-0.14
4	-0.10	-0.23	-0.14

Table 2: **Partners' relative contribution to the relationship between general PCA first dimension and clinical scales**. This table summarises the results of our partner-wise pairwise correlations, measuring the strength of association between the "general appraisal" of each individual partner, and clinical scales. For PAI-BOR totals and ECR-R-Anxiety ratings, partner 3 is prominent, the only partner whose association survived Bonferroni-Holmes correction for multiple comparisons. Partner 4 was somewhat associated with ECR-R-Anxiety ratings, and Partner 2 was the most prominent in terms of avoidance - however, these associations did not survive correction for multiple comparisons.

Item	PAI-BOR	ECR-R		PCA loading		
		Anxiety	Avoidance	dim.1	dim.2	dim.3
Likeable	-0.39^{**}	-0.48^{**}	-0.39^{**}	$+0.73$	$+0.35$	$+0.09$
Trustworthy	-0.42^{**}	-0.44^{**}	-0.30^{*}	$+0.73$	$+0.41$	$+0.20$
Made Happy	-0.28^{*}	-0.36^{*}	-0.23	$+0.66$	$+0.20$	-0.37
Made Sad	$+0.21$	$+0.35^{*}$	$+0.15$	-0.53	$+0.19$	$+0.54$
Irritating	$+0.28^{*}$	$+0.36^{*}$	$+0.29$	-0.78	$+0.17$	-0.09
Effort	-0.06	-0.09	$+0.02$	$+0.17$	$+0.73$	-0.25
Unpredictable	-0.03	-0.03	$+0.12$	-0.53	$+0.64$	-0.19
Unstable	$+0.14$	$+0.13$	$+0.05$	-0.72	$+0.38$	-0.21
Locus	-0.06	-0.02	$+0.02$	-0.07	$+0.43$	$+0.60$

Table 3: **General analysis of items marginalised across partners**. Columns specify the correlation coefficients between items marginalised across partners and PAI-BOR (col. 1), ECR-R-Anxiety (col. 2) and ECR-R-Avoidance (col. 3) questionnaire scores (** : $p < 0.001$;* : $p < 0.05$; p-values are Bonferroni-Holmes corrected for multiple comparisons). The three last columns report item loadings along the first two PCA dimensions. Scores along the first dimension are significantly associated with PAI-BOR and ECR-R sub-scales. The ECR-R-Avoidance subscale holds the strongest association with single items and, in turn, with the primary PCA dimension.

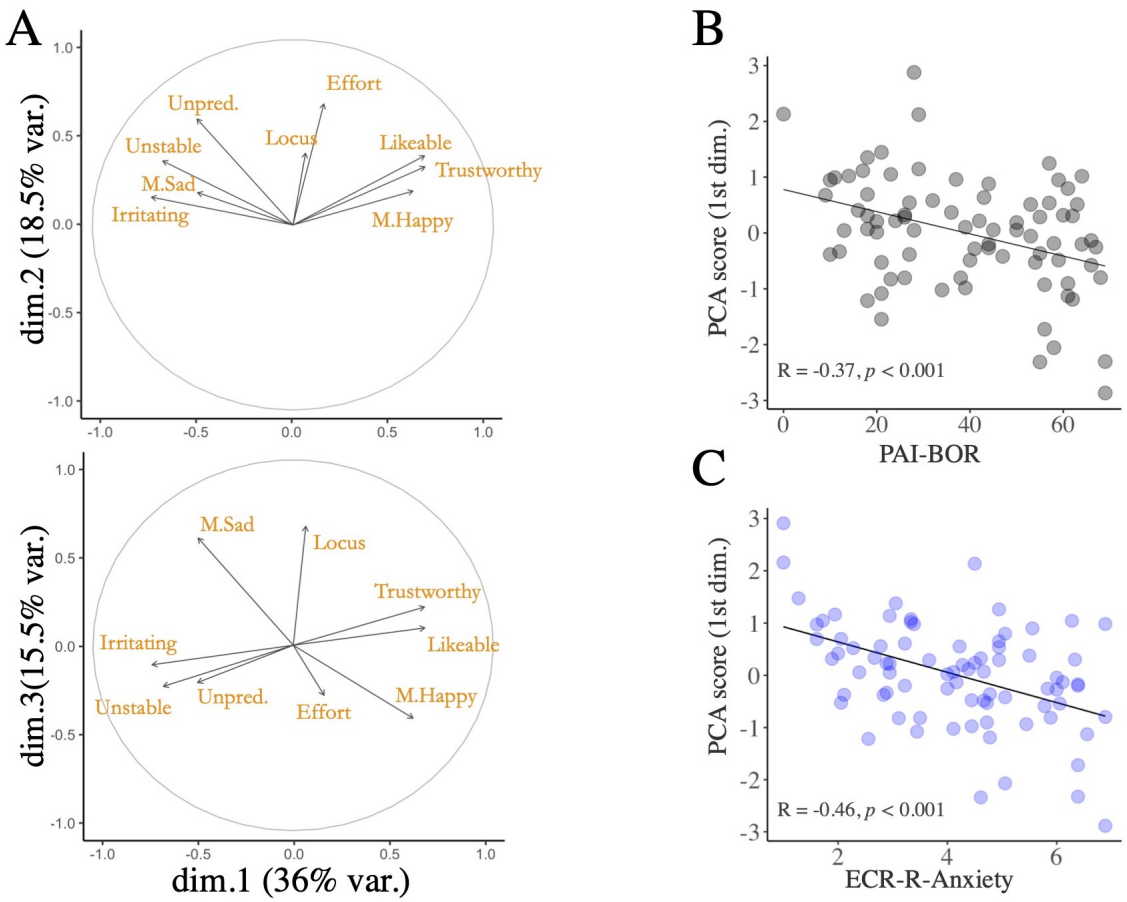

Figure 3: **PCA results**. Inset **A** illustrates the PCA loadings for marginalised items, juxtaposing the first and second dimensions (top) and first and third dimensions (bottom). The first PCA dimension (horizontal axis) reflects overall appreciation of the personality of *all* partners ("general appraisal"). Positive (negative) scores indicate a good (bad) general impression of dancing partners overall. The second dimension (top;vertical axis) reflects a tendency to give high or low scores. The third dimension (bottom;vertical axis) features good (and positive) contributions from Items *"M.Sad"* and *"Locus"*, a negative contribution from item *"M.Happy"* and rather neutral contributions from other items, so that it captures a tendency to feel responsible to have caused negative mood in partners. The first dimension was linked to both severity of borderline features as measured by PAI-BOR (inset **B**) and even more strongly to attachment Anxiety (ECR-R-Anxiety scores; inset **C**).

3.3. Partner-dependent analyses

Our key objective is to determine the role played by individual partners - the core of our experimental manipulation. To disentangle the extent to which specific partners related to our clinical scales, we projected the nine partner-specific items onto the "general appraisal" dimension we found, and related those with our clinical scales. Results are shown in table 2. Partner 3 was by-and-large the largest c. Notably, for attachment avoidance, partner 2 (showing on average the most positive affect) was by and large the most prominent contributor. This suggests that a negative experience with a partner showing more positive affect is most diagnostic for avoidant individuals, which marks a strong qualitative difference from the other dimensions. In table 3, we offer a more detailed summary of our PCA analyses, complemented by the pairwise correlations found between PAI-BOR and marginalised items.

3.4. Full sCCA analysis

We now move on to the exploratory analysis of the full questionnaire and clinical scales, the goal of which was to identify key partner-specific items linked with borderline symptomatology and its sub-domains. To obtain a quantitative link between dancing questionnaire and PAI-BOR subscales, we adopted a predictive approach using sparse canonical correlations analysis (sCCA; Witten et al., 2009; Witten & Tibshirani, 2009) in which, for simplicity, we only retained the first mode. We registered a good out-of sample performance for sCCA (median out-of-sample correlation coefficient = 0.32, min = 0.15, max = 0.41). When applied to the whole dataset, sCCA found a strong relationship between latent dimensions of the two sets of variables ($r = 0.59$, $p < 0.001$, $CI_{95\%} = [.41, .71]$), leaving the weight of dancing items nearly intact ($\lambda = 0.91$), and somewhat penalising questionnaire scales ($\lambda = 0.7$). The relative weights of items for partner 3 (biased in the negative range of expression) were most prominent (average of weights : 0.21; see Figure 4, inset D). Partner 1 followed with a lower contribution (0.14), and partners 2 and 4 were the least informative with even smaller average contributions (both 0.1). The importance of weights appears to follow closely the mood displayed by the partner over the course of the dance, in which the maximally negative mood observed weighs more than positive mood towards the ultimate judgement (recall that, due to the task design, subjects very frequently observed the extrema of the moods displayable by partners - i.e. their partner's best expression when they were successful in maintaining a good distance, and the worst when they were unsuccessful). Items "*Trustworthy*" and "*Made Happy*" for partner 3 had the largest sCCA weights across all questionnaire items (i.e. -0.35 for both items). The most important clinical scale, on the other hand, was the ECR-R-Anxiety sub-scale (scca weight: 0.73). In terms of PAI-BOR subscales, PAI-BOR-I (Identity problems) and PAI-BOR-S (Self-harm and Impulsivity) featured lower yet still sizeable contributions (0.50, and 0.46 respectively). We illustrate all sCCA results in Figure 4. In the Supplementary Material, we report the Bonferroni-Holmes corrected pair-wise correlations between dancing questionnaire and PAI and ECR-R sub-scales. Furthermore, We investigated whether there were partner-dependent items associated with our clinical questionnaires when considering healthy controls and patients separately. This was done to individuate features of the experience in the task that might be telling of symptom severity in the sub-clinical and clinical domains. We note that this was an exploratory (unplanned) analysis. Owing to the lower sample size caused by the split into controls ($N = 35$) and patients ($N = 45$) we used corrected pairwise correlations rather than sCCA for this analysis. Whilst we found no significant relationships when restricting ourselves to healthy controls' data , we found strong relationships between PAI-BOR and Trust in partners 1 ($r = -0.52, p_{adj} < 0.001, p < 0.001, CI_{adj95\%} = [-0.78, -0.09]$) and 3 ($r = -0.52, p_{adj} < 0.001, p < 0.001, CI_{adj95\%} = [-0.79, -0.08]$) when considering patients' data only. See Figure 5. Within subjects diagnosed with BPD, higher symptom severity is associated with decreased trust in partners 1 and 3 - those expressing a more negative range of affect. In healthy controls, attachment anxiety was associated with item *"Irritating"* concerning partner 1 ($r = 0.59, p_{adj} < 0.001, p < 0.001, CI_{adj95\%} = [0.11, 0.85]$) but not significantly associated with any item in patients (all corrected p-values > 0.45). Attachment avoidance was not significantly associated with any item in either the full dataset or when separating healthy controls from patients.

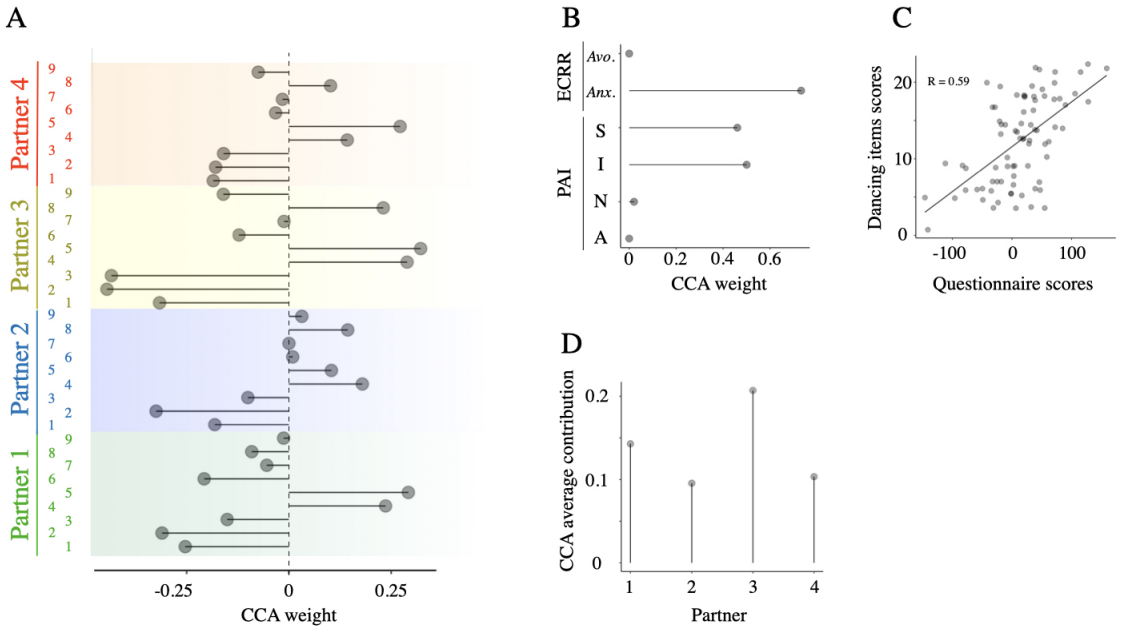

Figure 4: **Results of sCCA**. This figure summarises the results of our sCCA analyses. Inset **A** specifies the weights associated to each of the dancing questionnaire items. These are enumerated as described in the Methods, i.e. 1 : *"Likeable"*; 2 : *"Trustworthy"* ; 3 : *"Made Happy"* ; 4 : *"Made Sad"*; 5 : *"Irritating"* ; 6 : *"Effort"* ; 7 : *"Unpredictable"* ; 8 : *"Unstable"*; 9 : *"Locus"*. Consistent with PCA analyses over marginalised items (and pairwise correlations reported in Supplementary Materials) positive weights are weighed negatively, and negative ones positively. Inset **B** indicates the weights assigned to the questionnaire sub-scales which are the counterpart to the dancing questionnaire in our sCCA analysis. There is a salient contribution from attachment anxiety scores, followed by PAI-BOR-I (Identity problems) and PAI-BOR-S (Self-harm) with similar contributions. Attachment anxiety and PAI-BOR-I are strongly correlated in our sample (see Table 1). However the weight assigned to the former is larger. Inset **C** indicates the relationship between scores over the latent dimensions discovered. Finally, inset **D** includes a plot of the relative contribution of each individual partner in terms of sCCA weights (average of the absolute values of weights as depicted in A). The plot indicates a leading role of partner 3, followed by partner 1, which is in turn closely followed by partners 2 and 4 (the most uninformative).

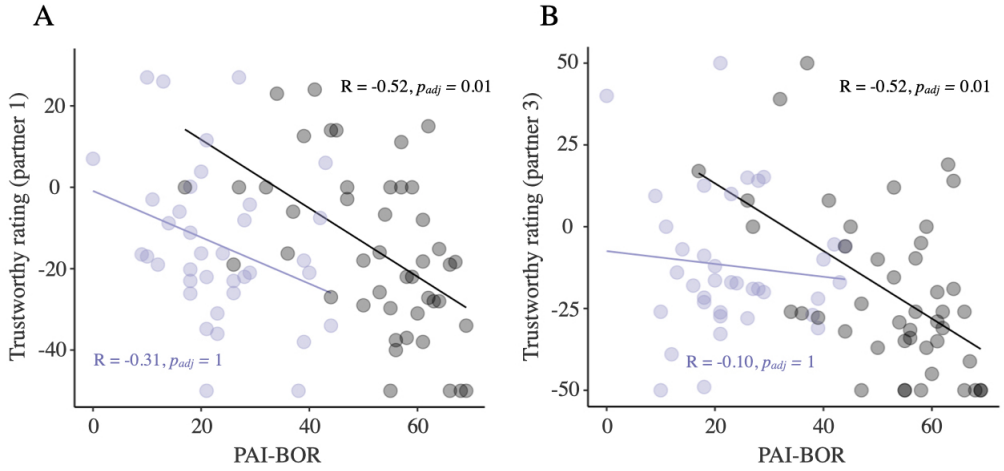

Figure 5: **Relationship between overall BPD symptom severity (PAI-BOR) and Trustworthiness of partners 1 (A) and 3 (B), in patients (black) and controls (blue)**. The relationship has the same direction in patients and controls but is stronger in patients, particularly for partner 3. Note that partners 1 and 3 are those capable of expressing the full range of negative facial expression.

3.5. Overt behaviour and proxemics

3.5.1. Reaction times

Reaction times did not vary significantly between conditions in which the partner was too close or too far from the subject. Accordingly, we pooled reaction times irrespective of subject and partner's relative position. Furthermore, our preliminary PCA analyses indicated a first dimension of variation conveying the general magnitude of reaction times, and a second dimension which reflected longer reaction times with partner 3 and shorter ones with the remaining partners (Figure 6). However, this latter explained only a small amount of variance (7.55%), which was largely accounted for by only the first dimension (81.4%). Thus, reaction times in our task did not differ across the two chosen modalities of relative subject-partner positions (being too close vs. too far) and did not exhibit meaningful variability across partners, so that they can simply be pooled together as a lumped, per-subject average reaction time. While here we report correlational results for the primary PCA dimension (as it follows more logically from our approach), we note that all relationships we describe are unvaried in effect size if we instead consider simple per-subject average reaction times. PAI-BOR scores correlated with the first PCA dimension ($r = 0.24$, $p = 0.04$, $CI_{95\%} = [0.01,0.42]$). The only sub-scale to correlate significantly with reaction time was PAI-BOR-N ($r = 0.29$, $p_{adj} = 0.03, p_u = 0.01$, $CI_{adj95\%} = [0.02, 0.53]$). Attachment anxiety (i.e., ECR-R-Anxiety) was particularly strongly associated with the primary reaction time dimension ($r = 0.38$, $p < 0.001$, $CI_{95\%} = [0.18,0.55]$).As an ulterior (unplanned) analysis, to compare effect sizes, we ran a simple a linear model in which both attachment anxiety and experience in negative relationships were used as predictors of reaction times. Attachment anxiety was found to be a much stronger predictor (ECR-R-Anxiety: $t = 2.18, p = 0.03$, PAI-BON: $t = 0.26, p = 0.79$). Finally, the degree of attachment anxiety predicted reaction times also when splitting the data by diagnosis (healthy controls: $r = 0.38$, $p = 0.03$, $CI_{95\%} = [0.05,0.63]$; patients: $r = 0.33$, $p = 0.04$, $CI_{95\%} = [0.04,0.57]$). In sum, attachment anxiety is a strong predictor of reaction times across all interactions, both when merging patients and controls and within the two groups separately.

3.5.2. Click distance from partner

Our analyses here consider the two separate conditions in which (i) subject and partner were too close, and (ii) too far (recall that in either case, the partner's mood is deteriorating). PCA of averaged click distances from partners when subject and partner were too close indicated a primary dimension of variation reflecting the general magnitude of click distance from partners (variance explained: 48.6%). The second dimension reflected larger distances clicked when with partners 1 and 3, and smaller when with partners 2 and 4. This second dimension explained a considerable amount of variance in the data (20.4%), so we kept this as a meaningful correlate. PAI-BOR scores were significantly correlated with the first dimension ($r = 0.26$, $p = 0.02$, $CI_{95\%} = [0.04,0.45]$). See Figure 6 so that higher PAI-BOR scores entailed larger click distances from partners both when partners were too close or far. PAI-BOR-A and PAI-BOR-S were the sub-scales underlying this relationship (correlations with the first dimension: PAI-BOR-A: $r = 0.29$, $p_{adj} = 0.02, p_u = 0.01$, $CI_{adj95\%} = [0.02,0.52]$; PAI-BOR-S: $r = 0.30$, $p_{adj} = 0.02, p_u = 0.01$, $CI_{adj95\%} = [0.02,0.52]$). The PCA dimensions for averaged click distances when subject and partner were too far apart again led to a primary dimension of variation reflecting the general magnitude of click distances from partners (variance explained: 54.1%), whilst the second dimension reflected larger distances clicked when with partners 1 and 3, and smaller when with partners 2 and 4 (variance explained: 21.2%). This condition however yielded no significant relationship of click distances with PAI BOR (p = 0.87) or with any of the sub-scales (all corrected p-values = 1).

3.5.3. Proportionate time spent with partners

A longer time spent with a given partner is indicative of a higher appreciation for such partner (see Table 2 in SMs). We calculated pairwise correlations between PAI-BOR scores and time spent with each partner - and found notable anti-correlations between PAI-BOR scores and time spent with partner 3 ($r = -0.21$, $p_u = 0.05$, $p_{adj} = 0.12$, $CI_{95\%} = [-0.45,0.05]$) and a similar trend for partner 4 ($r = 0.19$, $p_u = 0.08$, $p_{adj} = 0.12$, $CI_{95\%} = [-0.07,0.43]$). However, our study is under-powered for these sort of effect sizes - and these relationships do not survive correction for multiple comparisons.

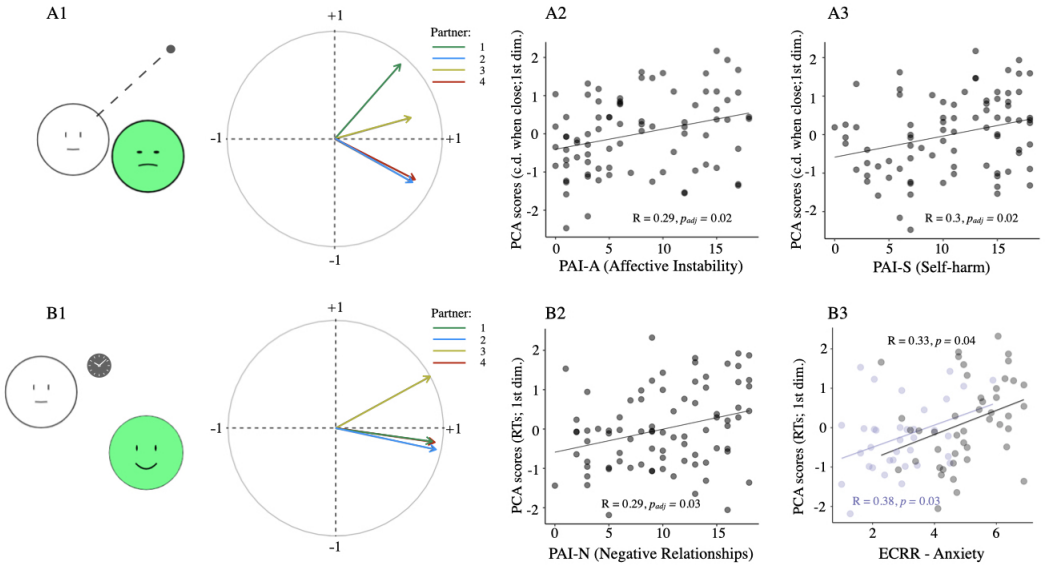

Figure 6: **Summary of proxemics analyses**. This figure illustrates the main results from our analyses on overt behaviour and proxemics. Insets **A1**,**A2**, and **A3** cover the click distance from the partner when the partner is too close - a proxy for intrusion in personal space (as suggested by the sketch on the left side of A1). Inset A1 summarises the principal components found - the first component (x-axis) which captures the general magnitude of distance clicked, and the second, which captures a differentiation in terms of partners (higher distances with partners 1 and 3, lower with 2 and 4). Only the first direction was significantly related to borderline severity, particularly the sub-domains of affective instability (PAI-BOR-A; **A2**) and self-harm (PAI-BOR-S; **A3**). The second row (**B1**,**B2**, **B3**) concerns reaction times and the relationships found between the first PCA dimension)which just summarises the general average time taken to make a move - without taking partners into account) and negative relationships (**B2**), and ECR-R-Anxiety (**B3**). This latter is the strongest relationship of the two, and holds across the whole dataset and when dividing subjects in healthy controls and patients.

4. Discussion

We have introduced a novel paradigm which takes us toward a second-person neuroscience (Schilbach et al., 2013). With this paradigm, we attempted to capture and describe the experience retained from a brief series of social interactions with previously unknown virtual partners. These differed in terms of biased emotional expression (partners 2 and 3), and unbiased expression with differing breadths (partners 1 and 4). Our analyses were based on behaviour during the task and post-task questionnaires. The questionnaire-based analyses determined the general and partner-specific items which were most strongly tied with the dimensional measures of BPD symptom severity and attachment style. In terms of the intrinsic relationship within our clinical scales of choice, sCCA analyses foregrounded the association between attachment anxiety and the identity problems sub-domain of borderline pathology. This confirms previous findings (Crawford et al., 2007) and aligns with the mentalizing perspective on personality disorder — in which identity problems (caused by and interplay between dispositional factors, ill-functioning child-primary caregiver relationships and/or trauma) can lead to patterns of anxious attachment and emotional dysregulation.

Our dimensionality reduction-based DTQ analyses revealed that scores along a primary dimension of appraisal of social interaction were negatively associated with attachment anxiety (ECR-R-Anxiety), avoidance (ECR-R-Avoidance) and borderline symptoms (PAI-BOR and subscales). Projecting partner-specific ratings on the latent dimensions revealed that more biasedly negative evaluations of partner 3 were those which best tied with BPD symptoms and our measure of attachment anxiety. Interestingly, items concerning partner 2 held the largest effect (albeit non-significantly post correction for multiple comparisons) in tying with attachment avoidance. This latter result is in line with previous evidence that avoidant individuals show a selective reduction of reported pleasantness for positive social stimuli (Vrtička et al., 2012). Taken

together, our dimensionality reduction-based results are consistent with a large body of work indicating that BPD sufferers hold negatively biased evaluations of others' in (Fertuck et al., 2013, 2019; Arntz & Veen, 2001; Barnow et al., 2009; Meyer et al., 2004; Nicol et al., 2013). The stronger relationship found for attachment anxiety (see Figure 3) was more unexpected, as this trait has been associated with valence-independent increased vigilance, rather than biased appraisal of others' behaviour. In a widely known paradigm bearing some analogy to ours in terms of the appraisal of morphing facial expressions (a modified version of the "morph movie" task: Niedenthal et al., 2000, 2001), Fraley et al. found that anxiously attached individuals were more sensitive to (i.e., were quicker to detect) variations in *all* emotional facial expression. Fraley et al. however reported smaller effect sizes for happy facial expressions, especially in terms of the transitions from neutral to happy emotional states. Our approach is of course different, as we do not measure the sensitivity to the onset/offset of an emotional expression - rather, we employ the DTQ to assess the integration of emotional expressions gathered over the course of all interactions. We just speculate that if transitions of facial expressions towards more negative states mattered more towards the ultimate appraisal (in anxiously attached individuals), the integration of such events would likely lead to the sort of relationship we observed here, with severity of attachment anxiety linked to a stronger negative appraisal overall. This extrapolation would be in line with previous literature indicating that anxiously attached individuals perceive more conflict in relationships and are hyper-vigilant about *negative* outcomes such as waning affection, or signs of potential withdrawal from their partners (Collins, 1996; Campbell et al., 2005) - and even experience more 'phantom vibrations' on their mobile phones when they "are concerned about something that [they] might get a call/message about" (Kruger & Djerf, 2016). In terms of mentalizing, the hyper-vigilance of anxiously attached individuals can be understood as an attempt to compensate for a reduced ability to mentalize with a propensity to engage in phenomenologically distinct yet ineffective hyper-mentalizing. From this perspective, the increased focus on negative affect in anxiously attached individuals is adaptive, since being rejected or abandoned is a threatening scenario to any social animal, and avoiding such an outcome warrants great effort. This makes our PCA dimension of generalised negative appraisal a good candidate for a parsimonious measure of individuals' sensitivity to negative social interactions, which could be a promising diagnostic tool for a specifc domain of mentalizing deficit. This asymmetry in sensitivity to positive and negative outcomes could also have important implications for psychotherapy.

Our more granular sCCA analyses revealed that items concerning Partner 3 were those which most strongly associated with PAI-BOR and ECR-R sub-scales overall. Notably, this was the partner who had the most negative range of affect. Partners 2 and 4 contributed substantially less to these associations while Partner 1 provided an intermediate contribution (Figure 4). The most prominent partner-specific relationships were those involving items *"Trustworthy"*, and *"Made Happy"* for Partner 3. The latter association is particularly interesting. It ties in with previous work suggesting that some forms of mental ill-health might be best characterized by a relatively impoverished — or possibly, 'unbiased' — way of updating affective beliefs and experiences of lack of locus of control and agency (Allport, 1955; Taylor & Brown, 1988; Sharot, 2011).

When we examined control and patient data separately, we found no significant relationships in healthy controls. This may be because of reduced power and thus requires replication in larger samples. In patients, however, we found a strong relationship between symptom severity and *"Trustworthy"* ratings concerning Partners 1 and 3 - those partners capable of manifesting the most negative affect. These results suggest that the appraisal of negative affect is a more precise predictor of symptom severity in patients.

When examining behaviour and proxemics directly, we found that severity of overall BPD symptoms as indicated by PAI-BOR was associated with slower reaction times, farther distancing on the next move when partners were too close, and (a trend toward) spending less time with the most-frowning partner (number 3). The former result aligns with work reporting slower reaction times in BPD patients (e.g., in facial trust appraisal: Fertuck et al., 2013, 2019). However, our finding underscores the role of experience in negative relationships (PAI-BOR-N) and attachment anxiety (ECR-R-Anxiety). In our study, longer reaction times could be signalling indecision - akin to inhibition, or uncertainty - a computational feature that may be linked with higher attachment anxiety. We found no previous instance of a link between larger reaction times and experience in negative relationships or attachment anxiety. This result should be replicated and

explored further.

Our observation concerning the relationship between BPD severity and larger distancing when partners were too close converges with previous studies which found that BPD patients have a larger preferred interpersonal distance (Fineberg et al., 2018a; Abdevali et al., 2021) and altered face processing in response to simulated intrusion in subjects' own personal space (Schienle et al., 2015). The regulation of personal space is thought to be a function of perceived emotional and physical threats, and varies with the level of intimacy and trust (Lloyd, 2009). In our study, the sub-scales involved in this relationship were PAI-BOR-A and PAI-BOR-S, suggesting that this effect might be related to a more vigorous reaction to a perceived intrusion - rather than serving the purpose of setting a comfortable interpersonal distance from the partner.

A number of limitations concerning the present must be pointed out. First, with this being a novel task, some analyses were exploratory - and need replicated. While we observed our item-wise results both through sCCA and Bonferroni-Holmes corrected correlations, we can not prove to be able to disambiguate between the individual effect of single items on the relationships we found - since items are intrinsically correlated with each other - and permuting correlated items does not alter results in a significant way. Thus, while it is interesting that item *"Made Happy"* might be especially diagnostic, we must await further replications of our study to gain confidence that this item deserves a special merit. Second, our analyses of overt behaviour were approached conservatively, by first decomposing data into relevant dimensions and only establishing relationships of clinical questionnaires with the quantities thus found a posteriori. Future task developments should make it possible to attempt more cohesive explanations of behaviour and microscopic (motor) decision-making, perhaps through the use of computational models - the use of which could be very insightful. Third, the task may too short, which entails that we can not reliably measure some aspects of behaviour - for instance, our results on proportion of time spent with Partners 3 and 4 indicate only a trend. Failure to observe a more robust effect may be due to the fact that subjects had only 14 minutes to play, and were made to play with all partners at least once. Future iterations of our study using longer versions of this task might offer more variability in the proportion of time spent with partners, which could strengthen the relationship observed. Finally, it would be interesting to add a socially goal-directed component to the task - such that the social interactions included are not ephemeral but are needed to establish trust - for instance to reach a decision about whose advice to trust in a final decision that must be made after the dances. Alternatively, one could provide a more ecological meaning to the act of touching - such that when the subject and partner's avatars touch, subjects become vulnerable to them, for instance vulnerable to the avatar either giving or taking away money from a final bonus given for the participant's time.

Taken together, our results support the notion that our newly developed task (and the approach that it operationlizes) can uncover and quantify known and unknown aspects of healthy and ill-functioning social appraisal. Our task operationalized partner personality in a straightforward way – by manipulating the range of facial expression – and our results speak for a strongly asymmetric weighing of negatively valued expressions. We know of no previous paradigms which have studied the impact of co-occurring positive and negatively valued stimuli when appraising a novel acquaintance, especially in a clinical population known to be vulnerable to compromized attribution of intentions. We have provided robust evidence that higher ratings in terms of attachment disturbances and borderline symptoms tie with a negatively biased appraisal of novel social interactions, and added to this result by observing a somewhat novel, powerful explanatory role for attachment anxiety, also in terms of measures of overt behaviour and proxemics (e.g. slower reaction times). By focusing on borderline sub-scales, we could pinpoint facets of borderline symptomatology (identity problems and impulsivity/self harm) which varied best with negative appraisal using sCCA. Future work should of course replicate our initial findings - and further refine and expand our paradigm – building upon which we may be able to obtain, at some point in the future, useful diagnostic tools or therapeutic aids.

References

Abdevali, M., Mazaheri, M. A., Besharat, M. A., Zabihzadeh, A., & Green, J. D. (2021). Borderline personality disorder and larger comfortable interpersonal distance in close relationships. *Personality and Individual Differences*, *182*, 111067.

Agrawal, H. R., Gunderson, J., Holmes, B. M., & Lyons-Ruth, K. (2004). Attachment studies with borderline patients: A review. *Harvard review of psychiatry*, *12*, 94–104.

Aldunate, N., & González-Ibáñez, R. (2017). An integrated review of emoticons in computer-mediated communication. *Frontiers in psychology*, *7*, 2061.

Allport, G. W. (1955). *Becoming: Basic considerations for a psychology of personality* volume 20. Yale University Press.

Anupama, V., Bhola, P., Thirthalli, J., & Mehta, U. M. (2018). Pattern of social cognition deficits in individuals with borderline personality disorder. *Asian journal of psychiatry*, *33*, 105–112.

APA (2013). Diagnostic and statistical manual of mental disorders. *Fifth Edition*, *21*.

Arntz, A., & Veen, G. (2001). Evaluations of others by borderline patients. *The Journal of nervous and mental disease*, *189*, 513–521.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and social psychology bulletin*, *29*, 819–833.

Barakova, E., Gillessen, J., & Feijs, L. (2009). Social training of autistic children with interactive intelligent agents. *Journal of Integrative Neuroscience*, *8*, 23–34.

Barnow, S., Stopsack, M., Grabe, H. J., Meinke, C., Spitzer, C., Kronmüller, K., & Sieswerda, S. (2009). Interpersonal evaluation bias in borderline personality disorder. *Behaviour research and therapy*, *47*, 359–365.

Behrens, T. E., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. (2008). Associative learning of social value. *Nature*, *456*, 245–249.

Berenson, K. R., Dochat, C., Martin, C. G., Yang, X., Rafaeli, E., & Downey, G. (2018). Identification of mental states and interpersonal functioning in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *9*, 172.

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, *16*, 74.

Bertsch, K., Hillmann, K., & Herpertz, S. C. (2018). Behavioral and neurobiological correlates of disturbed emotion processing in borderline personality disorder. *Psychopathology*, *51*, 76–82.

Campbell, L., Simpson, J. A., Boldry, J., & Kashy, D. A. (2005). Perceptions of conflict and support in romantic relationships: the role of attachment anxiety. *Journal of personality and social psychology*, *88*, 510.

Collins, N. L. (1996). Working models of attachment: Implications for explanation, emotion, and behavior. *Journal of personality and social psychology*, *71*, 810.

Crawford, T. N., John Livesley, W., Jang, K. L., Shaver, P. R., Cohen, P., & Ganiban, J. (2007). Insecure attachment and personality disorder: A twin study of adults. *European Journal of Personality: Published for the European Association of Personality Psychology*, *21*, 191–208.

De Panfilis, C., Riva, P., Preti, E., Cabrino, C., & Marchesi, C. (2015). When social inclusion is not enough: Implicit expectations of extreme inclusion in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *6*, 301.

Domsalla, M., Koppe, G., Niedtfeld, I., Vollstädt-Klein, S., Schmahl, C., Bohus, M., & Lis, S. (2014). Cerebral processing of social rejection in patients with borderline personality disorder. *Social cognitive and affective neuroscience*, *9*, 1789–1797.

Fertuck, E., Jekal, A., Song, I., Wyman, B., Morris, M., Wilson, S., Brodsky, B., & Stanley, B. (2009). Enhanced 'reading the mind in the eyes' in borderline personality disorder compared to healthy controls. *Psychological medicine*, *39*, 1979–1988.

Fertuck, E. A., Grinband, J., Mann, J. J., Hirsch, J., Ochsner, K., Pilkonis, P., Erbe, J., & Stanley, B. (2019). Trustworthiness appraisal deficits in borderline personality disorder are associated with prefrontal cortex, not amygdala, impairment. *NeuroImage: Clinical*, *21*, 101616.

Fertuck, E. A., Grinband, J., & Stanley, B. (2013). Facial trust appraisal negatively biased in borderline personality disorder. *Psychiatry research*, *207*, 195–202.

Fineberg, S. K., Leavitt, J., Landry, C. D., Neustadter, E. S., Lesser, R. E., Stahl, D. S., Deutsch-Link, S., & Corlett, P. R. (2018a). Individuals with borderline personality disorder show larger preferred social distance in live dyadic interactions. *Psychiatry research*, *260*, 384–390.

Fineberg, S. K., Leavitt, J., Stahl, D. S., Kronemer, S., Landry, C. D., Alexander-Bloch, A., Hunt, L. T., & Corlett, P. R. (2018b). Differential valuation and learning from social and nonsocial cues in borderline personality disorder. *Biological psychiatry*, *84*, 838–845.

Fineberg, S. K., Stahl, D. S., & Corlett, P. R. (2017). Computational psychiatry in borderline personality disorder. *Current behavioral neuroscience reports*, *4*, 31–40.

First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B., & Benjamin, L. S. (1997). *Structured clinical interview for DSM-IV® axis ii personality disorders SCID-II*. American Psychiatric Pub.

Fossati, A., Maffei, C., Bagnato, M., Donati, D., Namia, C., & Novella, L. (1999). Latent structure analysis of dsm-iv borderline personality disorder criteria. *Comprehensive Psychiatry*, *40*, 72–79.

Fraley, R. C., Niedenthal, P. M., Marks, M., Brumbaugh, C., & Vicary, A. (2006). Adult attachment and the perception of emotional expressions: Probing the hyperactivating strategies underlying anxious attachment. *Journal of personality*, *74*, 1163–1190.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of self-report measures of adult attachment. *Journal of personality and social psychology*, *78*, 350.

Frick, C., Lang, S., Kotchoubey, B., Sieswerda, S., Dinu-Biringer, R., Berger, M., Veser, S., Essig, M., & Barnow, S. (2012). Hypersensitivity in borderline personality disorder during mindreading. *PloS one*, *7*, e41650.

Gluschkoff, K., Jokela, M., & Rosenström, T. (2021). General psychopathology factor and borderline personality disorder: Evidence for substantial overlap from two nationally representative surveys of us adults. *Personality Disorders: Theory, Research, and Treatment*, *12*, 86.

Gunderson, J. G. (2007). Disturbed relationships as a phenotype for borderline personality disorder. *American Journal of Psychiatry*, *164*, 1637–1640.

Gunderson, J. G. (2009). *Borderline personality disorder: A clinical guide*. American Psychiatric Pub.

Gunderson, J. G., Zanarini, M. C., Choi-Kain, L. W., Mitchell, K. S., Jang, K. L., & Hudson, J. I. (2011). Family study of borderline personality disorder and its sectors of psychopathology. *Archives of General Psychiatry*, *68*, 753–762.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*, 243–259.

Henco, L., Diaconescu, A. O., Lahnakoski, J. M., Brandi, M.-L., Hörmann, S., Hennings, J., Hasan, A., Papazova, I., Strube, W., Bolis, D. et al. (2020). Aberrant computational mechanisms of social learning and decision-making in schizophrenia and borderline personality disorder. *PLoS computational biology*, *16*, e1008162.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Johansen, M., Karterud, S., Pedersen, G., Gude, T., & Falkum, E. (2004). An investigation of the prototype validity of the borderline dsm-iv construct. *Acta Psychiatrica Scandinavica*, *109*, 289–298.

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*, 806–810.

Kruger, D. J., & Djerf, J. M. (2016). High ringxiety: Attachment anxiety predicts experiences of phantom cell phone ringing. *Cyberpsychology, behavior, and social networking*, *19*, 56–59.

Laming, D. R. J. (1968). Information theory of choice-reaction times., .

Leichsenring, F., Leibing, E., Kruse, J., New, A. S., & Leweke, F. (2011). Borderline personality disorder. *The Lancet*, *377*, 74–84.

Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M., & Bohus, M. (2004). Borderline personality disorder. *The Lancet*, *364*, 453–461.

Lloyd, D. M. (2009). The space between us: A neurophilosophical framework for the investigation of human interpersonal space. *Neuroscience & Biobehavioral Reviews*, *33*, 297–304.

Lowyck, B., Luyten, P., Vanwalleghem, D., Vermote, R., Mayes, L. C., & Crowley, M. J. (2016). What's in a face? mentalizing in borderline personality disorder based on dynamically changing facial expressions. *Personality Disorders: Theory, Research, and Treatment*, *7*, 72.

Maier, W., Lichtermann, D., Klingler, T., Heun, R., & Hallmayer, J. (1992). Prevalences of personality disorders (dsm-iii-r) in the community. *Journal of personality disorders*, *6*, 187.

Mathiak, K., & Weber, R. (2006). Toward brain correlates of natural behavior: fmri during violent video games. *Human brain mapping*, *27*, 948–956.

Matzke, B., Herpertz, S. C., Berger, C., Fleischer, M., & Domes, G. (2014). Facial reactions during emotion recognition in borderline personality disorder: a facial electromyography study. *Psychopathology*, *47*, 101–110.

McCall, C. (2015). Mapping social interactions: the science of proxemics. In *Social Behavior from Rodents to Humans* (pp. 295–308). Springer.

Meyer, B., Pilkonis, P. A., & Beevers, C. G. (2004). What's in a (neutral) face? personality disorders, attachment styles, and the appraisal of ambiguous social cues. *Journal of personality disorders*, *18*, 320–336.

Michael, J., Chennells, M., Nolte, T., Ooi, J., Griem, J., Network, M. D. R., Christensen, W., Feigenbaum, J., King-Casas, B., Fonagy, P. et al. (2021). Probing commitment in individuals with borderline personality disorder. *Journal of Psychiatric Research*, *137*, 335–341.

Mitchell, A. E., Dickens, G. L., & Picchioni, M. M. (2014). Facial emotion processing in borderline personality disorder: a systematic review and meta-analysis. *Neuropsychology review*, *24*, 166–184.

Moran, P., Leese, M., Lee, T., Walters, P., Thornicroft, G., & Mann, A. (2003). Standardised assessment of personality–abbreviated scale (sapas): preliminary validation of a brief screen for personality disorder. *The British Journal of Psychiatry*, *183*, 228–232.

Morey, L. C. (1991). *Personality assessment inventory*. Psychological Assessment Resources Odessa, FL.

Nicol, K., Pope, M., Sprengelmeyer, R., Young, A. W., & Hall, J. (2013). Social judgement in borderline personality disorder. *PLoS One*, *8*, e73440.

Niedenthal, P. M., Brauer, M., Halberstadt, J. B., & Innes-Ker, Å. H. (2001). When did her smile drop? facial mimicry and the influences of emotional state on the detection of change in emotional expression. *Cognition & Emotion*, *15*, 853–864.

Niedenthal, P. M., Halberstadt, J. B., Margolin, J., & Innes-Ker, Å. H. (2000). Emotional state and the detection of change in facial expression of emotion. *European journal of social psychology*, *30*, 211–222.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*, 559–572.

Renneberg, B., Heyn, K., Gebhard, R., & Bachmann, S. (2005). Facial expression of emotions in borderline personality disorder and depression. *Journal of behavior therapy and experimental psychiatry*, *36*, 183–196.

Ritzl, A., Csukly, G., Balázs, K., & Égerházi, A. (2018). Facial emotion recognition deficits and alexithymia in borderline, narcissistic, and histrionic personality disorders. *Psychiatry research*, *270*, 154–159.

Schienle, A., Wabnegger, A., Schöngassner, F., & Leutgeb, V. (2015). Effects of personal space intrusion in affective contexts: an fmri investigation with women suffering from borderline personality disorder. *Social cognitive and affective neuroscience*, *10*, 1424–1428.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience 1. *Behavioral and brain sciences*, *36*, 393–414.

Sevgi, M., Diaconescu, A. O., Henco, L., Tittgemeyer, M., & Schilbach, L. (2020). Social bayes: using bayesian modeling to study autistic trait–related differences in social cognition. *Biological Psychiatry*, *87*, 185–193.

Sharot, T. (2011). The optimism bias. *Current biology*, *21*, R941–R945.

Skodol, A. E., Gunderson, J. G., Pfohl, B., Widiger, T. A., Livesley, W. J., & Siever, L. J. (2002). The borderline diagnosis 1: psychopathology, comorbidity, and personaltity structure. *Biological psychiatry*, *51*, 936–950.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, *103*, 193.

Unoka, Z., Seres, I., Aspan, N., Bódi, N., & Kéri, S. (2009). Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. *Journal of personality disorders*, *23*, 399–409.

Vrtička, P., Sander, D., & Vuilleumier, P. (2012). Influence of adult attachment style on the perception of social and non-social emotional scenes. *Journal of Social and Personal Relationships*, *29*, 530–544.

WHO (2004). The international statistical classification of diseases and health related problems icd-10: Tenth revision, . *1*.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, *10*, 515–534.

Witten, D. M., & Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, *8*.

Xiang, T., Ray, D., Lohrenz, T., Dayan, P., & Montague, P. R. (2012). Computational phenotyping of two-person interactions reveals differential neural response to depth-of-thought. *PLoS computational biology*, *8*.

Yuasa, M., Saito, K., & Mukawa, N. (2006). Emoticons convey emotions without cognition of faces: an fmri study. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 1565–1570).