


Do You See What I See? Researcher-Participant Agreement on Single-Item Measures of Emotion Regulation Behaviors in Borderline Personality Disorder

Assessment
1–9
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10731911211044216
journals.sagepub.com/home/asm


Nicole E. Stumpp¹ , Matthew W. Southward¹ ,
and Shannon Sauer-Zavala¹

Abstract

Researchers use ecological momentary assessment (EMA) to study a range of behaviors related to psychopathology. However, it is unclear whether brief measures of coping behaviors accurately capture the intended responses. In this secondary analysis of a single-case experimental design, eight individuals with borderline personality disorder ($M_{age} = 21.57$, 63% female, 63% Asian American) completed daily diary entries for 12 weeks, along with hourly EMA entries on 2 days. Participants provided qualitative descriptions of their behaviors and classified them into one of five functional categories. Independent researchers also classified each qualitative description into the same categories. Overall, agreement between participants and researchers was low, Krippendorff's $\alpha = .47$, 95% confidence interval [0.43, 0.52]. The type of emotion experienced, researcher confidence, and word count of responses affected agreement. Generating items that capture the breadth of possible behaviors, are brief enough for frequent administration, and are consistently understood by participants is an important continued challenge in EMA research.

Keywords

ecological momentary assessment, borderline personality disorder, emotion-driven behaviors, coping behaviors, daily diary

Borderline personality disorder (BPD) is a severe psychiatric disorder characterized by a pervasive pattern of affective and behavioral instability (American Psychiatric Association, 2013). However, tracking this instability over time has proven difficult. Researchers have found discrepancies between retrospective self-reports and real-time assessments of mood, symptoms, traits, and behaviors (Fahrenberg et al., 2007; Shiffman et al., 2008). People with BPD more accurately recall their experiences on days without significant mood shifts and struggle to retrospectively recreate days with such shifts (Solhan et al., 2009). Additionally, people with BPD tend to overestimate the intensity of negative moods and underestimate the intensity of positive moods in the context of post-hoc self-reports (Santangelo et al., 2014). People with untreated BPD may also be less accurate at identifying and naming their emotions compared with healthy controls and people with BPD being treated with Dialectical Behavior Therapy (Ebner-Priemer et al., 2007; Ebner-Priemer et al., 2008). These difficulties may be particularly exacerbated during periods of heightened psychological distress. Taken together, these findings suggest that assessing mood fluctuations in people with BPD using retrospective self-report may provide different information than ratings captured in vivo.

One sampling technique researchers have used to address these difficulties is ecological momentary assessment (EMA). In EMA designs, participants use a personal device (e.g., cell phone) to report phenomena of interest as they occur in real time. EMA reduces the recall bias associated with retrospective self-report, strengthens ecological validity, and is better equipped to capture affective variability and instability in those with BPD (Shiffman et al., 2008). Ratings of emotional intensity in particular may translate readily to EMA batteries (Santangelo et al., 2014). For example, existing validated questionnaires of emotional experiences (e.g., positive and negative affective schedule (PANAS; Watson et al., 1988) have been abbreviated for frequent real-time administration (e.g., Jacobson et al., 2020). There is also evidence that using single items representing discrete emotions (i.e., indicating the extent to which a person is experiencing sadness or anxiety) is a valid

¹University of Kentucky, Lexington, KY, USA

Corresponding Author:

Nicole E. Stumpp, University of Kentucky, 343 Waller Avenue, Suite 303, Lexington, KY 40504, USA.
Email: nicole.stumpp@uky.edu

approach to capture affective experiences (Harmon-Jones et al., 2016; Jacobson et al., 2020).

Two studies have assessed affective instability characteristic of BPD using the PANAS in samples of individuals with BPD (Hepp et al., 2020) and individuals with a range of psychopathology (Ringwald et al., 2020). Importantly, these studies also assessed behavioral responses to emotions and interpersonal stressors at each EMA beep and found that more variability in the types of behavioral responses reflected maladaptive emotion dysregulation (Ringwald et al., 2020) and that interpersonal stressors were associated with greater negative affect (Hepp et al., 2020). Although these studies contribute to the relatively sparse literature characterizing affective instability using EMA methods, neither study collected qualitative data regarding how participants responded behaviorally to interpersonal stressors. Furthermore, the studies that have assessed behavioral responses to stressful stimuli using EMA methods among participants with BPD typically only assess engagement in a single maladaptive behavior, such as rumination or suppression (e.g., Chapman et al., 2009; Yaroslavsky et al., 2019).

Thus, it is unclear whether brief measures can accurately capture the full range of coping behaviors used by people with BPD. When measuring emotion regulation, researchers have typically applied one of two approaches: (1) using multi-item scales to assess the use of relatively fewer regulation strategies (e.g., Brockman et al., 2017; Medland et al., 2020) or (2) using single items to capture a broader range of regulatory behaviors (e.g., Heiy & Cheavens, 2014; Southward & Cheavens, 2020; Southward et al., 2019). Unfortunately, multi-item scales cannot account for the full nature of a person's emotion regulation repertoires and single-item scales often demonstrate relatively lower content validity and reliability. In two systematic reviews of studies utilizing EMA methods, many researchers approached measurement development by simply selecting items from cross-sectional measures and adapting the wording to fit the study's timeframe (Griffin et al., 2020; Trull & Ebner-Priemer, 2020). Furthermore, although certain EMA measures of emotion regulation behaviors may demonstrate internal validity and reliability, one assumption underlying all these measures is that participants are interpreting the EMA response choices in line with the researchers' intentions. However, without direct participant feedback describing their behaviors, it is unclear whether the behavioral response participants select from an EMA item to describe their regulatory behavior aligns with that intended by the researchers.

Current Study

In a recent trial examining the impact of a brief treatment on the frequency of maladaptive behavioral coping in people

with BPD, Sauer-Zavala et al. (2020) asked participants to categorize their daily emotion regulation behaviors using five relatively broad categories (i.e., emotional avoidance, emotional savoring, engage in impulsive behavior, problem solving, acceptance). Participants also provided qualitative descriptions of each behavior. In the current study, we conducted a secondary data analysis of the agreement between participants and researchers regarding participants' reports and descriptions of their emotion regulation behaviors from Sauer-Zavala et al. (2020), Southward et al. (2020) and Cardona et al. (2020). Specifically, our research team independently categorized participants' qualitative responses as one of five behavioral response options presented to participants during the parent study and examined agreement with participants' own categorizations of their behaviors. We hypothesized that researchers and participants would demonstrate acceptable agreement regarding these classifications. We then explored several potential moderators (i.e., type of emotion experienced, emotion intensity, researcher confidence, severity of BPD symptoms, treatment responder status) that may affect the observed level of agreement. Given the exploratory nature of these moderators, no specific hypotheses were made.

Method

Participants

Participants included eight individuals with BPD ($M_{age} = 21.57$, $SD = 3.05$). The majority of the sample was female (63%), Asian American (63%), non-Hispanic (88%), and not taking psychotropic medication (88%). Individuals were included if they met the following criteria: (1) *Diagnostic and statistical manual of mental disorders—5th edition (DSM-5)* (American Psychiatric Association, 2013) diagnosis of BPD, (2) willingness to maintain their current dose of psychotropic medication throughout the duration of the study, (3) willingness to abstain from additional psychosocial treatment for the duration of the study, (4) fluency in English, and (5) access to a personal smartphone. Participants were excluded if they endorsed conditions that necessitated prioritization of alternative treatment. Specifically, exclusion criteria consisted of the following: (1) current manic episode, schizophrenia, schizoaffective disorder, or other organic mental disorder, (2) current acute suicidal risk, and (3) current or recent (within 3 months) history of substance dependence.

Participants were recruited through local treatment sites, online postings, and direct contact with individuals with BPD who had previously participated in a nontreatment study conducted by the research team. Potential participants completed a phone screen and those likely to be eligible were invited to the research clinic to complete a semistructured, diagnostic assessment to confirm eligibility. Of the

10 individuals that completed this baseline assessment, eight were eligible to participate in this study. All enrolled participants ($N = 8$) completed all study procedures. Details about study flow and full participant demographics can be viewed in Sauer-Zavala et al. (2020).

Study Design

The primary study from which these data are derived (Sauer-Zavala et al., 2020) utilized a single-case experimental design (Barlow et al., 2009). The study consisted of three phases: baseline, intervention, and follow-up. Specifically, participants were randomly assigned to a baseline period of either 2 or 4 weeks followed by an intervention phase consisting of 4 weekly sessions of the Countering Emotional Behaviors module of the Unified Protocol (Barlow et al., 2018). Finally, participants completed a 4-week assessment-only follow-up phase. Participants completed daily diary entries throughout all study phases, as well as hourly entries on two randomly chosen days (one during the baseline phase and one during the last week of the intervention phase). Notifications were sent via text message or email (based on participant preference) to participants' smartphones to gather information regarding their current emotional experiences, the triggers that prompted these emotions, and their behavioral responses to them. All study procedures were approved by the local university institutional review board and participants provided their informed consent prior to engaging in study-related activities.

Measures

Diagnostic Assessment. The Structured Clinical Interview for *DSM-IV* Axis II Disorders–Borderline Personality Disorder Module (First et al., 1997) was used to identify the presence/absence of BPD according to *DSM-IV-TR* criteria, which is identical to *DSM-5* criteria, in the current study. The Structured Clinical Interview for *DSM-IV* Axis II Disorders–Borderline Personality Disorder Module has demonstrated good psychometric properties (Ryder et al., 2007) and strong interrater reliability in prior research ($\kappa = .91$; Lobbestael et al., 2011). Additionally, specific modules from the Anxiety Disorders Interview Schedule for *DSM-5* (Brown & Barlow, 2014) were used to assess study exclusion criteria. Advanced doctoral students administered these instruments to participants at their initial study visit and demonstrated excellent agreement regarding study eligibility ($\kappa = 1.00$; Sauer-Zavala et al., 2020).

BPD Symptom Severity. Participants completed the self-report version of the Zanarini Rating Scale for Borderline Personality Disorder (ZAN-BPD; Zanarini et al., 2015) weekly throughout all study phases. The ZAN-BPD is a nine-item measure designed to assess the severity of each of

the *DSM-IV-TR* (American Psychiatric Association, 2000) criteria for BPD. Participants rate the degree to which they have experienced each symptom in the prior week using a 0 to 4 scale with unique anchors for each item. ZAN-BPD items have demonstrated good internal consistency (Cronbach's $\alpha = .84$) and 7- to 10-day test–retest reliability ($r = .66$) in previous research (Zanarini et al., 2015). In the current study, ZAN-BPD items demonstrated acceptable average internal consistency across all weeks (Cronbach's $\alpha = .78$; range: .56-.91)¹ and good 1-week test–retest reliability ($r = .67$).

Daily Diary and EMA Assessments of Emotional Experiences

Daily diary. Each day of the study, participants were prompted to report on the characteristics of their emotional experiences. Participants were first asked to identify which emotion(s), if any, they had experienced since their last diary report from a list of anger, anxiety, guilt/shame, and sadness. For each emotion identified, participants rated its intensity from 1 (*no intensity at all*) to 5 (*greatest possible intensity*). Finally, they were asked to select which of five behavioral responses they used in response to each emotion. The response options presented were chosen to broadly capture the range of behavioral responses individuals with BPD may engage in to regulate their emotions. Although each behavioral response was presented in lay English to participants in a way tailored to the emotion after which it was presented, responses represented five general domains: problem solving, acceptance, impulsive responding, emotional avoidance, and emotional savoring. For example, if a participant reported feeling anxious, emotional savoring was described as “dug in to the feeling (e.g., repeated checking, extra preparation for an event, cleaned, sought reassurance),” whereas if they reported feeling sad, emotional savoring was described as “dug in to the feeling (e.g., isolated myself, cried, listened to sad music, watched a sad movie).” Full texts of these response options are available in Southward et al. (2020).

Between each rating, participants were prompted to provide qualitative descriptions of their emotional experiences and responses. After identifying an emotion experienced, participants were asked to describe in their own words what triggered that emotion. Similarly, after rating the intensity of the emotion, but before identifying a behavioral response, participants were asked to describe what they did in response to the emotion. Participants were not limited in how much, or how little, they could write in response to each prompt.

EMA. On two randomly selected days, one during the baseline phase and one during the last week of the intervention phase, participants completed nearly identical reports of their emotional experiences every hour for 12 consecutive hours. Participants were first asked to rate the intensity of their current emotional experience from 0 (*no intensity*

at all) to 100 (*greatest intensity*). If participants selected 0, the survey ended. If participants indicated an intensity of 1 or higher, they were asked to select which of five emotions best described their current experience: anger, sadness, anxiety, guilt/shame, or joy/happiness. They were then asked to provide qualitative free response descriptions of what triggered this emotion and what they were doing in response to the emotion. Finally, participants were asked to select which of five behavioral categories best characterized their free response description of their behavior, using the same categories tailored to the emotion experienced provided in the daily diaries. However, in contrast to the daily diaries, participants could only describe one emotional experience. This was done to reduce participant burden and because we assumed participants would experience fewer emotions from 1 hour to the next compared with the course of the entire day.

Coding behavioral responses. The first author (NES) and a research assistant were trained by the author of the parent study (SSZ) to identify the type of behavior(s) referenced by each EMA item using hypothetical examples and group consensus discussion to resolve discrepancies. The first author then read each participant's qualitative descriptions of their behavioral responses and classified each description as one of the five categories of behavioral responses that were provided to participants: problem-solving, acceptance, impulsive responding, emotional avoidance, or emotional savoring. A research assistant completed the same procedures for a randomly selected 20% of the total qualitative descriptions provided ($n = 194$). For negative emotions (anxiety, anger, sadness, guilt/shame), responses categorized as problem solving or acceptance were considered within the overarching label of "adaptive," whereas responses categorized as impulsive responding, emotional avoidance, and emotional savoring were labeled as "maladaptive." When joy/happiness was endorsed in the EMA entries, we considered savoring to be adaptive, rather than maladaptive, because this strategy involves up-regulating the positive emotion experienced. Because participant ratings of emotional intensity differed between the daily diaries (1-5 scale) and EMA entries (0-100 scale), daily intensity ratings were recoded to match the EMA scale (i.e., 1 = 0, 2 = 25, 3 = 50, 4 = 75, 5 = 100).

Researcher confidence. For each response coded, researchers also indicated their degree of confidence in each coding. Confidence was rated on a 3-point Likert-type scale from 1 (*not very confident*) to 3 (*very confident*).

Analytic Method

We first examined descriptive statistics of the number and frequency of participants' responses. We then tested trends

in participants' rates of responding over the study. To determine if the average number of daily responses provided by participants differed as a function of study phase, we ran a repeated measures analysis of variance in SPSS Version 27 using phase as a within-person factor. We followed up this analysis of variance with post-hoc tests to examine differences among phases. We were powered to detect large effects, $f_s > .63$, $F_s > 3.73$.

We then merged participants' data from the daily diaries and hourly EMA. We calculated Krippendorff's alpha to assess agreement both between researchers and between researchers' and participants' categorizations of their behaviors using the KALPHA macro (Hayes & Krippendorff, 2007) in SPSS Version 27. Krippendorff's α , which is bounded by $[-1.00, 1.00]$ quantifies the degree of agreement beyond chance between n researchers on nominal, ordinal, interval, or ratio levels of data with any degree of missingness. Krippendorff (2004) has suggested that $\alpha_s \geq .80$ indicate relatively reliable variables and α_s from .67 to .80 indicate tentative reliability. The KALPHA macro allows for the application of a bootstrapping algorithm to calculate 95% confidence intervals (CIs) around point estimates of alpha by randomly resampling subsets of the original data set. We generated 5,000 bootstrap resamples when calculating α .

Next, we explored whether agreement between researchers and participants varied according to aspects of the study design, participant responses, participant background characteristics, and researcher perceptions. Because responses were nested within participants, and because we coded agreement dichotomously (*no agreement* = 0, *agreement* = 1), we used hierarchical logistic mixed modeling with the *glmer* function in the *lme4* package (Bates et al., 2015) in R Version 4.0.3 (R Core Team, 2020) to model predictors of agreement. We modeled random intercepts² and applied bound optimization by quadratic approximation with 10 integration points in all models for consistency and greater comparability. In separate models, we explored if study design characteristics such as study phase, daily versus hourly responses, and number of words used in participants' qualitative responses were associated with the likelihood of agreement between participants and researchers on any given response. We then tested if aspects of participant responses (i.e., emotion type and intensity) were related to the likelihood of agreement in separate hierarchical logistic mixed models.

Given the relatively small number of participants, we calculated Spearman's signed-rank correlations and used Wilcoxon signed rank tests to examine the relations between participant characteristics and agreement. These nonparametric tests are more appropriate for small samples because they do not assume variables are normally distributed. We used the *cor.test* function (with the Spearman specifier) in R to calculate Spearman's correlation between baseline BPD

features and average agreement for each participant. We were powered to detect large effects, $\psi_s > .76$. We then used the *wilcox.test* function in R to run a Wilcoxon signed-rank test comparing the average proportion of agreement with each participant between those who responded to the intervention to those who were considered nonresponders. Intervention responders were defined as those participants who demonstrated reductions in the frequency with which they used emotionally avoidant behaviors in response to strong emotions during the follow-up phase that were non-overlapping with emotionally avoidant behaviors reported during the baseline phase. We were powered to detect large effects, $d_s > 2.55$.

Finally, to explore researcher characteristics, we again used hierarchical logistic mixed modeling to predict the likelihood of agreement. Because researcher confidence may be influenced by the number of words participants used to describe their behaviors, we also tested the relation between researcher confidence and word count, and the interaction of these variables to predict likelihood of agreement. We plotted the resulting interaction using the *sjPlot* package (Lüdtke, 2021) in R. All code can be found at <https://doi.org/10.17605/osf.io/xbfap>.

Results

Descriptive Statistics

Participants provided 917 total responses: 774 daily responses and 143 hourly responses. Participants failed to provide a response for 52 entries. Each participant provided an average of 114.63 total responses ($SD = 35.68$), with 96.75 ($SD = 30.77$) daily responses and 17.88 ($SD = 8.49$) hourly responses. Participants reported using problem solving most frequently ($n = 256$; 26.4%), followed by emotional avoidance ($n = 236$; 24.4%), emotional savoring ($n = 203$; 20.9%), acceptance ($n = 178$; 18.4%), and impulsive responding ($n = 38$; 3.9%). Participants reported feeling anxious most frequently when pushing feelings away ($n = 104$; 44.1%). When reporting emotional avoidance, they reported feeling anger most frequently ($n = 63$; 31.0%). When reporting impulsive responding, participants most frequently reported feeling anxious ($n = 13$; 34.2%). They reported feeling anxious most frequently when categorizing their behavior as problem solving ($n = 133$; 52.0%). They reported feeling anxiety most frequently when using emotional savoring as well ($n = 104$; 58.4%). Participants reported feeling happiness/joy on 52 (5.4%) entries, 24 (3.9%) of which were reported by treatment responders. Participants used 7.14 ($SD = 8.17$) words on average to describe their behaviors at each entry. There were significant differences in the average number of daily responses provided in each study phase, $F(2, 14) = 11.91$, $p < .01$. Participants provided more daily responses on average during both the baseline phase ($M = 42.13$, $SD =$

16.30), $M_{\text{difference}} = 26.75$, $SD_{\text{difference}} = 18.11$, $p < .01$, 95% CI [11.61, 41.89], and intervention phase ($M = 39.25$, $SD = 17.79$), $M_{\text{difference}} = 23.88$, $SD_{\text{difference}} = 18.19$, $p < .01$, 95% CI [8.67, 39.08], compared with the follow-up phase ($M = 15.38$, $SD = 4.84$).

Agreement in EMA Ratings

Nineteen responses were not coded by the research team due to blank qualitative responses or responses that did not describe behaviors (e.g., “I cannot fall asleep,” “fear,” “picked wrong button”). Two independent researchers from the research team demonstrated excellent agreement with each other when categorizing the subset of participants’ qualitative responses into the five behavioral categories above, Krippendorff’s $\alpha = .90$, 95% CI [0.86, 0.95], 85.6% raw agreement. As with participants, researchers rated problem solving most frequently ($n = 284$; 29.3%), followed by emotional avoidance ($n = 246$; 25.4%), emotional savoring ($n = 200$; 20.6%), acceptance ($n = 121$; 12.5%), and impulsive responding ($n = 48$; 5.0%). However, researchers demonstrated relatively low agreement with participants when categorizing qualitative responses into five behavioral response categories, Krippendorff’s $\alpha = .47$, 95% CI [0.43, 0.52], 60.0% raw agreement. Agreement between researchers and participants remained relatively low when classifying participant’s behavioral responses as either adaptive or maladaptive, Krippendorff’s $\alpha = .57$, 95% CI [0.51, 0.62], 78.4% raw agreement. Participants tended to categorize their behaviors as adaptive more frequently than the researchers: researchers coded 43.1% ($n = 418$) of behavioral responses as adaptive, whereas participants coded 48.5% ($n = 470$) of responses as adaptive.

Correlates of Agreement

We then explored whether certain characteristics of the study design, participant responses, participant background characteristics, and researcher perceptions were related to the likelihood of researcher and participant agreement in a series of hierarchical logistic mixed models. Study phase (i.e., baseline, intervention, or follow-up) was not associated with the likelihood of agreement, $B = -0.09$, $SE = 0.10$, $p = .37$, 95% CI [-0.28, 0.11]. Likewise, likelihood of agreement did not differ significantly when comparing hourly EMA responses to daily entries, $B = 0.35$, $SE = 0.19$, $p = .07$, 95% CI [-0.03, 0.72]. Furthermore, the number of words participants used to describe their behaviors was unrelated to the likelihood of agreement with researchers, $B = 0.01$, $SE = 0.01$, $p = .51$, 95% CI [-0.01, 0.03].

Emotion Type and Intensity. By contrast, the type of emotion experienced was associated with agreement likelihood. Compared with behavioral responses when experiencing happiness (38.3% raw agreement), researchers and participants

Table 1. Individual Treatment Response and Match Rates.

Participant	Treatment response	Matches n (%)	Adaptive ratings (%)	Coded adaptive ratings ^a (%)	Krippendorff's α [95% CI]	Total diary responses ^b
1	Non-responder	71 (78.0)	30 (33.0)	32 (35.2)	.65 [0.51, 0.77]	91
2	Non-responder	56 (70.0)	61 (76.3)	60 (75.0)	.57 [0.43, 0.71]	80
3	Responder	97 (53.3)	84 (46.2)	50 (27.5)	.37 [0.26, 0.46]	182
4	Responder	74 (52.9)	69 (49.3)	63 (45.0)	.38 [0.27, 0.49]	140
5	Responder	43 (52.9)	52 (59.8)	38 (43.7)	.31 [0.17, 0.45]	87
6	Responder	59 (54.6)	68 (63.0)	70 (64.8)	.38 [0.26, 0.51]	108
7	Responder	49 (61.3)	53 (66.3)	49 (61.3)	.47 [0.31, 0.60]	80
8	Non-responder	90 (69.2)	53 (40.8)	56 (43.1)	.58 [0.48, 0.69]	130

^aIndicates the percentage of adaptive ratings made by the researcher team. ^bIndicates the total number of diary entries the researcher team coded.

were more likely to agree when categorizing responses to sadness, $B = 1.17$, $SE = 0.35$, $p < .01$, 95% CI [0.49, 1.88], 63.6% raw agreement; anxiety, $B = 1.10$, $SE = 0.33$, $p < .01$, 95% CI [0.46, 1.76], 61.7% raw agreement; anger, $B = 1.06$, $SE = 0.35$, $p < .01$, 95% CI [0.38, 1.76], 60.2% raw agreement; and guilt/shame, $B = 0.82$, $SE = 0.39$, $p = .04$, 95% CI [0.06, 1.60], 55.3% raw agreement. However, the intensity of the emotional experience, regardless of the type of emotion, was unrelated to the likelihood of agreement, $B = -0.002$, $SE = 0.003$, $p = .39$, 95% CI [-0.008, 0.003].

Participant Characteristics. Baseline BPD symptom severity was not associated with average raw agreement with each participant, $\rho = 0.05$, $p = .91$. However, agreement did differ as a function of participants' treatment response, $W = 15$, $p < .04$, 95% CI [0.08, 0.29]. Participants who responded to the intervention demonstrated lower raw agreement (53.9%) on average with the research team than participants who did not respond to the intervention (72.1%). See Table 1 for a detailed breakdown of agreement by participants.

Coding Characteristics. On average, researcher confidence scores were high ($M = 2.78$, $SD = 0.54$) with 753 scores of 3 on a 3-point scale.³ Greater researcher confidence in categorizations was associated with a greater likelihood of agreement with participants, $B = 1.92$, $SE = 0.21$, $p < .01$, 95% CI [1.53, 2.36]. By contrast, researcher confidence was associated with participants using fewer words to describe their responses, $B = -0.009$, $SE = 0.002$, $p < .01$, 95% CI [-0.015, -0.005]. The interaction of word count and researcher confidence was significantly associated with agreement likelihood, $B = -0.05$, $SE = 0.02$, $p < .01$, 95% CI [-0.08, -0.01] (Figure 1). Agreement tended to be higher when participants used more words to describe their behavior, particularly when researcher confidence was low to medium. However, when participants used fewer words to describe their behavior, agreement was more strongly and positively related to researcher confidence.

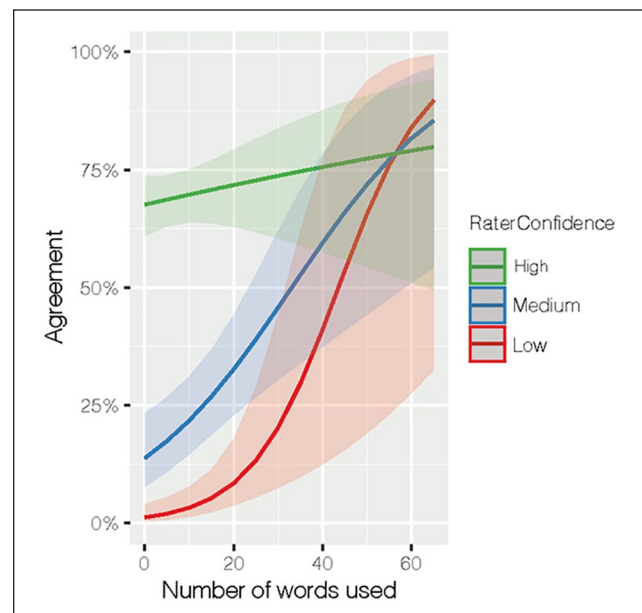


Figure 1. Likelihood of agreement between participants and researchers as a function of word count and researcher confidence.

Discussion

In this secondary data analysis, we observed relatively low rates of agreement between participants and independent researchers when classifying emotional behaviors into five broad categories, as well as two summary categories (i.e., adaptive vs. maladaptive). Agreement was more likely among participants who did not respond to the parent study's intervention and when participants experienced negatively valenced emotions. Researcher confidence and number of words also predicted agreement, such that agreement was higher when participants used more words or when researchers reported greater confidence. However, study phase, frequency of reporting, emotional intensity, and BPD features were each unrelated to agreement. We discuss each of these findings and their implication for EMA researchers in turn.

Our primary hypothesis that researchers and participants would demonstrate acceptable agreement when categorizing behavioral responses to emotions was not supported. This pattern of results may suggest (1) that the categories used in the behavioral response item did not accurately distinguish specific behaviors or provide broad enough coverage of behavioral options, (2) that participants did not understand the varieties of behaviors intended to be captured by the categories, (3) that participants' brief descriptions did not provide enough detail for researchers to accurately categorize each behavior, or (4) that participants require more psychoeducation to accurately categorize their behaviors. Although five relatively broad response categories were included, in line with many EMA studies of emotion regulation behaviors (Brockman et al., 2017; Medland et al., 2020), it is possible that there was too much conceptual overlap among these categories. For instance, examples provided to participants of emotional avoidance and impulsive responding both included substance use and self-injury with the implication that emotional avoidance was relatively more intentional whereas impulsive responding was more spontaneous. Relatedly, the response options provided excluded other relevant behaviors (e.g., reappraisal, social support) that may have been easier for participants and researchers to identify.

Second, participants may have misunderstood the different types of behaviors that could be captured by each response option. In the parent study, participants did not receive training in categorizing their behaviors given that the primary goal was to test a novel treatment and we wanted to minimize the chances that the EMA component would influence behavior. Alternatively, participants may not have provided enough or the appropriate detail for researchers to accurately categorize their behaviors. We found that number of words participants used interacted with researcher confidence in predicting agreement. This interaction suggests that the use of more words was associated with better agreement regardless of researcher confidence, but the use of fewer words was only related to better agreement when researcher confidence was high. High researcher confidence in the presence of few words likely indicates participants used key words to describe their behavior that made categorizing the behavior easier for researchers and participants.

Although response options that provide greater conceptual coverage may improve agreement between participants and researchers, it is also important to consider that agreement was lower when using five categories than with two. Of course, providing fewer categories may also artificially inflate agreement due to chance. This effect can be seen when comparing the increase in raw agreement when moving from five to two categories (+18.4%) to the relatively smaller increase in kappa values from five to two categories (+0.10). We encourage researchers to use measures of agreement, such as Krippendorff's α , that minimize this

effect, while recognizing that providing as few categories as necessary to fully capture the range of behaviors may contribute to higher agreement and lessen participant burden.

It is also possible that people with BPD require extended psychoeducation to gain insight into their coping strategies. One would expect insight into emotion regulation behaviors to improve as participants are learning about emotions and emotion regulation. However, the lack of increased agreement between researchers and participants as the study progressed (i.e., study phase was not associated with agreement) suggests that more psychoeducation may be necessary for participants to accurately categorize their behavioral responses. Additionally, because intervention response was determined by self-reported decreases in emotionally avoidant behaviors, the lower agreement observed between researchers and intervention responders may be due to participants believing their behaviors are more adaptive than independent observers would rate. We encourage researchers to include validation checks during EMA studies to ensure participants understand the material as researchers intend and to track symptomatology throughout treatment.

Other than researcher confidence and word count, the only factors that were associated with better agreement were negatively valenced emotions, compared with happiness/joy, and intervention nonresponse, compared with intervention response. Participants who did not respond to the intervention in the parent study likely also reported more frequent negative emotions than those who did respond. These results suggest that more frequent behavioral responses to negative emotions may generate better agreement. Participants who experience negative emotions more frequently may be more familiar with their responses to those emotions and thus more clearly describe them for researchers. This population may benefit from learning and focusing on positive emotions and emotion regulation strategies in treatment. Alternatively, this finding may be a result of the behavioral responses assessed. Although the lay description of each response was adapted to be appropriate for happiness/joy, the five behavioral options may not have fully captured the likely range of responses participants would use when experiencing happiness/joy. It is noteworthy that in one EMA study comparing the regulation of positive and negative emotions, almost half of the strategies used to regulate positive emotions did not have a conceptually similar strategy to regulate negative emotions (Heiy & Cheavens, 2014). Thus, it is possible that providing a combination of similar and distinct response options for positive and negative emotions may enhance the understanding of categories and agreement between participants and researchers.

Continued research is needed to clarify the best methods for capturing coping behaviors as they occur in real time. For example, it is possible that specific features of BPD (i.e., lack of emotional clarity, misunderstanding one's behavioral responses) make it particularly difficult for

individuals with this condition to accurately classify their behaviors using brief EMA items. Future research should compare participant accuracy in a variety of emotional disorders (and with individuals without clinically significant psychopathology) to better understand the influence symptom severity and disorder type have on match rates. Another possibility is that collapsing the broad range of coping behaviors into five EMA-friendly categories may not provide the most utility for participants to conceptualize their behaviors. Generating items that capture the breadth of possible behaviors, are brief enough for frequent administration in an EMA context, and are consistently understood by participants is an important challenge for researchers. Providing an accessible way for participants to conceptualize their behavioral responses so researchers can categorize these responses as adaptive or maladaptive at the data analytic stage may be more advantageous than asking participants to group their behaviors. Finally, this study is limited by its small sample size given that the data are drawn from a single-case experimental design; thus, it is difficult to draw conclusions regarding whether results will generalize to the full range of psychopathology.

Despite these limitations, the present study is among the first to test the validity of an EMA item designed to characterize emotion-driven behaviors in BPD. The results can help future researchers understand how to accurately capture behavioral responses to emotional experiences while providing an accessible way for participants to conceptualize these experiences. These findings can further clarify how BPD may impact agreement about behavioral responses to emotions when using brief EMA items.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Nicole E. Stumpp  <https://orcid.org/0000-0001-9871-6469>

Matthew W. Southward  <https://orcid.org/0000-0002-5888-2769>

Notes

1. Solutions for McDonald's ω did not converge, most likely due to the relatively small sample size.
2. Models with random slopes demonstrated worse fit as judged by Akaike Information Criteria, so we only report models with random intercepts.
3. Example responses when researcher confidence was high include "I took a long nap" (rated as "pushing the feeling away"

by both researchers and participants) and "Felt they belittled me—I confronted them" (rated as "digging in" by researchers and as "problem-solving" by participants), to which participant categorization did and did not match researcher coding, respectively. Example responses when researcher confidence was low include "Internalized it" (rated as "digging in" by both researchers and participants), and "I cried again, and I tried to ignore it" (rated as "digging in" by researchers and "pushing the feeling away" by participants), to which participant categorization did and did not match researcher coding, respectively.

References

- American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (Text Rev.). <https://doi.org/10.1002/9780470479216.corpsy0271>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single Case Experimental Designs: Strategies for Studying Behavior Change* (3rd edition). Boston, MA: Pearson.
- Barlow, D. H., Sauer-Zavala, S., Farchione, T. J., Murray Latin, H., Ellard, K. K., Bullis, J. R., Bentley, K. H., Boettcher, H. T., & Cassiello-Robbins, C. (2018). *Unified Protocol for the Transdiagnostic Treatment of Emotional Disorders: Patient Workbook* (2nd ed.). New York, NY: Oxford University Press
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. <https://doi.org/10.18637/jss.v067.i01>
- Brockman, R., Ciarrochi, J., Parker, P., & Kashdan, T. (2017). Emotion regulation strategies in daily life: Mindfulness, cognitive reappraisal and emotion suppression. *Cognitive Behaviour Therapy*, *46*(2), 91-113. <https://doi.org/10.1080/16506073.2016.1218926>
- Brown, T.A., & Barlow, D. H. (2014). *Anxiety and Related Disorders Interview Schedule for DSM-5 (ADIS-5L): Client Interview Schedule* (Lifetime version). New York, NY: Oxford University Press.
- Cardona, N. D., Southward, M. W., Furbish, K., Comeau, A., & Sauer-Zavala, S. (2020). Nomothetic and idiographic patterns of behavioral responses to emotions in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, *12*(4), 354-364. <https://doi.org/10.1037/per0000465>
- Chapman, A. L., Rosenthal, M. Z., & Leung, D. W. (2009). Emotion suppression in borderline personality disorder: An experience sampling study. *Journal of Personality Disorders*, *23*(1), 29-47. <https://doi.org/10.1521/pedi.2009.23.1.29>
- Ebner-Priemer, U. W., Kuo, J., Schlotz, W., Kleindienst, N., Rosenthal, M. Z., Detterer, L., Linehan, M. M., & Bohus, M. (2008). Distress and affective dysregulation in patients with borderline personality disorder: A psychophysiological ambulatory monitoring study. *Journal of Nervous and Mental Disease*, *196*(4), 314-320. <https://doi.org/10.1097/NMD.0b013e31816a493f>
- Ebner-Priemer, U. W., Welch, S. S., Grossman, P., Reisch, T., Linehan, M. M., & Bohus, M. (2007). Psychophysiological ambulatory assessment of affective dysregulation in border-

- line personality disorder. *Psychiatry Research*, 150(3), 265-275. <https://doi.org/10.1016/j.psychres.2006.04.014>
- Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory assessment: Monitoring behavior in daily life settings. *European Journal of Psychological Assessment*, 23(4), 206-213. <https://doi.org/10.1027/1015-5759.23.4.206>
- 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*. Washington, D.C.: American Psychiatric Press.
- Griffin, S. A., Freeman, L. K., & Trull, T. J. (2020). *Measuring impulsivity in daily life: A systematic review and recommendations for future research*. PsyArXiv Preprints. <https://doi.org/10.31234/osf.io/mfdp6>
- Harmon-Jones, C., Bastian, B., & Harmon-Jones, E. (2016). The Discrete Emotions Questionnaire: A new tool for measuring state self-reported emotions. *PLOS ONE*, 11(8), e0159915. <https://doi.org/10.1371/journal.pone.0159915>
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1, 77-89.
- Heiy, J., & Cheavens, J. (2014). Back to basics: A naturalistic assessment of the experience and regulation of emotion. *Emotion*, 14(5), 878-891. <https://doi.org/10.1037/a0037231>
- Hepp, J., Lane, S. P., Carpenter, R. W., & Trull, T. J. (2020). Linking daily-life interpersonal stressors and health problems via affective reactivity in borderline personality and depressive disorders. *Psychosomatic Medicine*, 82(1), 90-98. <https://doi.org/10.1097/PSY.0000000000000728>
- Jacobson, N. C., Evey, K. J., Wright, A. G. C., & Newman, M. G. (2020). *Integration of discrete and global structures of affect across three large samples: Specific emotions within-persons and global affect between-persons*. PsyArXiv Preprints. <https://doi.org/10.31234/osf.io/gb5up>
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.
- Lobbestael, J., Leurgans, M., & Arntz, A. (2011). Inter-researcher reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). *Clinical Psychology & Psychotherapy*, 18(1), 75-79. <https://doi.org/10.1002/cpp.693>
- Lüdtke, D. (2021). *sjPlot: Data visualization for statistics in social science* (Version 2.8.7) [Computer software]. CRAN. <https://CRAN.R-project.org/package=sjPlot>
- Medland, H., De France, K., Hollenstein, T., Mussoff, D., & Koval, P. (2020). Regulating emotion systems in everyday life: Reliability and validity of the RESS-EMA Scale. *European Journal of Psychological Assessment*, 36(3), 437-446. <https://doi.org/10.1027/1015-5759/a000595>
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Australia, URL <https://www.R-project.org/>.
- Ringwald, W. R., Hopwood, C. J., Pilkonis, P. A., & Wright, A. G. C. (2020). Dynamic features of affect and interpersonal behavior in relation to general and specific personality pathology. *Personality Disorders: Theory, Research, and Treatment*. Advance online publication. <https://doi.org/10.1037/per0000469>
- Ryder, A. G., Costa, P. T., & Bagby, R. M. (2007). Evaluation of the SCID-II Personality Disorder Traits for DSM-IV: Coherence, discrimination, relations with general personality traits, and functional impairment. *Journal of Personality Disorders*, 21(6), 626-637. <https://doi.org/10.1521/pedi.2007.21.6.626>
- Santangelo, P., Bohus, M., & Ebner-Priemer, U. W. (2014). Ecological momentary assessment in borderline personality disorder: A review of recent findings and methodological challenges. *Journal of Personality Disorders*, 28(4), 555-576. https://doi.org/10.1521/pedi_2012_26_067
- Sauer-Zavala, S., Cassiello-Robbins, C., Woods, B. K., Curreri, A., Wilner Tirpak, J., & Rassaby, M. (2020). Countering emotional behaviors in the treatment of borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment*, 11(5), 328-338. <https://doi.org/10.1037/per0000379>
- Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4(1), 1-32. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091415>
- Solhan, M. B., Trull, T. J., Jahng, S., & Wood, P. K. (2009). Clinical assessment of affective instability: Comparing EMA indices, questionnaire reports, and retrospective recall. *Psychological Assessment*, 21(3), 425-436. <https://doi.org/10.1037/a0016869>
- Southward, M. W., & Cheavens, J. S. (2020). More (of the right strategies) is better: Disaggregating the naturalistic between- and within-person structure and effects of emotion regulation strategies. *Cognition and Emotion*, 34(8), 1729-1736. <https://doi.org/10.1080/02699931.2020.1797637>
- Southward, M. W., Heiy, J. E., & Cheavens, J. S. (2019). Emotions as context: Do the naturalistic effects of emotion regulation strategies depend on the regulated emotion? *Journal of Social and Clinical Psychology*, 38(6), 451-474. <https://doi.org/10.1521/jscp.2019.38.6.451>
- Southward, M. W., Semcho, S. A., Stumpp, N. E., MacLean, D. L., & Sauer-Zavala, S. (2020). A day in the life of borderline personality disorder: A preliminary analysis of within-day emotion generation and regulation. *Journal of Psychopathology and Behavioral Assessment*, 42(4), 702-713. <https://doi.org/10.1007/s10862-020-09836-1>
- Trull, T. J., & Ebner-Priemer, U. W. (2020). Ambulatory assessment in psychopathology research: A review of recommended reporting guidelines and current practices. *Journal of Abnormal Psychology*, 129(1), 56-63. <https://doi.org/10.1037/abn0000473>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063-1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Yaroslavsky, I., Napolitano, S. C., & France, C. M. (2019). Ruminative responses to interpersonal precipitants mediate borderline personality disorder features' effects on distress reactivity and recovery in daily life. *Journal of Clinical Psychology*, 75(12), 2188-2209. <https://doi.org/10.1002/jclp.22839>
- Zanarini, M. C., Weingeroff, J. L., Frankenburg, F. R., & Fitzmaurice, G. M. (2015). Development of the self-report version of the Zanarini Rating Scale for borderline personality disorder. *Personality and Mental Health*, 9(4), 243-249. <https://doi.org/10.1002/pmh.1302>