Refinement of the Reflective Function Questionnaire for Youth (RFQY) Scale B using Item Response Theory

Carla Sharp (1,2)*, Lynne Steinberg (1)*, Veronica McLaren (1), Stuart Weir (3), Carolyn Ha (4),

and Peter Fonagy (5)

(1) University of Houston

(2) University of the Free State, Center for Development Support

(3) Baylor College of Medicine

(4) Katy Psychological Services

(5) University College London

Corresponding author:
Carla Sharp, Ph.D.
Department of Psychology
University of Houston
126 Heyne Building
csharp2@uh.edu

*Sharp and Steinberg contributed equally to the manuscript.

Abstract

We conducted item response theory (IRT) analyses to refine the Reflective Function Questionnaire for Youth (RFQY) Scale B. Data from a non-clinical sample of young people (n = 737; ages 18-25) was used to derive a shortened version of the RFQY. Results were replicated in a clinical sample of inpatient adolescents (n = 467; ages 12-17), resulting in a five-item measure, thereafter named the RFQY-5. The RFQY-5 item set was then scrutinized for construct validity against the original 23-item RFQY item set in a randomly selected sample of 100 inpatient adolescents not included in the IRT replication, and 186 healthy adolescents drawn from the community. Results showed that the RFQY-5 performed similarly as the long version in terms of associations with criterion variables, and outperformed the longer version in discriminating between inpatient and community-dwelling adolescents who differed in their levels of borderline traits. The study provides evidence in support of the use of the RFQY-5 in research and clinical settings.

Refinement of the Reflective Function Questionnaire for Youth (RFQY-Scale B) using Item Response Theory

Mentalization-based theory is a social-cognitive theory of personality developed by Fonagy and colleagues (Bateman & Fonagy, 2004; Fonagy, 1991; Fonagy, Gergely, Jurist, & Target, 2002). As a theory it is wide in its scope, as it incorporates the disciplines of attachment, philosophy of mind, developmental science, cognitive science, and neuroscience. One way of operationalizing mentalization is through the construct of reflective function (RF). RF is a multi-component construct defined as the capacity to imagine and recognize mental states in self and others in the context of an attachment relationship (Fonagy, Steele, Moran, Steele, & Higgitt, 1991). RF is considered “multi-component”, because the construct includes reflection on the thoughts and feelings about self and others, and is seen to occur both implicitly (without conscious awareness) and explicitly (with conscious awareness). In essence, this capacity allows an individual to not only react to another person’s behavior, but to his or her perception of what underlies that behavior in relation to an individual’s own intentions and behaviors (Fonagy & Target, 1997). According to mentalization-based theory and practice, optimal RF reflect openness and uncertainty towards the mental states of self and others, which foster curiosity and flexibility in interpersonal interactions.

Empirical research on the construct of RF has been increasing steadily over the last two decades and the construct has shown high relevance and usefulness for parent-child attachment, psychopathology and psychotherapy research (Katznelson, 2014). While deficits in RF have been linked to various psychiatric problems including autism, depression, psychosis, PTSD, eating disorders, substance abuse and psychopathy (for a review see Katznelson, 2014; Pajulo et al., 2012; Schechter et al., 2006; Toth, Rogosch, & Cicchetti, 2008), it is perhaps most closely associated with borderline personality disorder (BPD) in which the concept of RF first emerged (Fonagy, 1991) and was popularized (Fonagy et al., 2016). Consistent with theoretical positions, studies in adult patients with BPD have emphasized the importance of promoting RF abilities through psychotherapy (Bouchard et al., 2008; Levy et al., 2006). Indeed, increasing RF capacity is seen as the main focus of one of the major evidence-based psychotherapies for BPD, Mentalization-based Therapy (Bateman & Fonagy, 2016). Given the importance of RF in treatment settings especially for personality disorder, timely and valid assessments of RF are important.

Several methods have been developed for the assessment of RF. The predominant approach has been the Adult Reflective Function Scale (ARFS) on the Adult Attachment Interview (AAI; George, Kaplan, & Main, 1985). However, the time and energy demanded by the ARFS renders this measure somewhat difficult and cumbersome to administer and limits its use as a clinical scale (Hill, Levy, Meehan, & Reynoso, 2007; Meehan, Levy, Reynoso, Hill, & Clarkin, 2009). In reaction, a 46-item self-report measure of RF was developed by Fonagy and Ghinai (2008) and named the Reflective Function Questionnaire (RFQ). In parallel, the adult RFQ was adapted for use in adolescents given the importance of the adolescent and young adult developmental periods for social-cognitive maturity (Sharp & Fonagy, 2015) and named the Reflective Function Questionnaire for Youth (RFQY; Sharp et al., 2009). Moreover, there has been increased consensus that borderline pathology is a valid construct in adolescents (see Chanen, 2015; Chanen & Kaess, 2012; Sharp, 2016; Sharp & Fonagy, 2015; Sharp & Romero, 2007 for reviews). Thus, by measuring RF in adolescents and young adults, clinicians have a greater chance of early detection of key processes associated with borderline pathology and implementing timely evidenced-based interventions to improve the prognosis of youths diagnosed with borderline pathology (Fonagy et al., 2015).

Mirroring the adult RFQ (Fonagy and Ghiani, 2008)), the RFQY has two subscales. These subscales have no substantive relation to the multiple components of the RF concept (e.g. self vs. other or implicit vs. explicit). Instead, as originally designed in the adult version, the two subscales reflect two different scoring approaches. Scale A includes items where optimal RF responses are located at the mid-point of a 6-point Likert-type scale, with low RF responses located near *either* endpoint, while Scale B includes items where RF responses are scored from low to high using a 6-point Likert-type scale. The purpose of mid-point scoring in Scale A reflects the acknowledgement that too much RF may not be optimal depending on the wording of the item. For instance, if the statement "I always know what others are thinking" is endorsed with response option 6 (the highest on the Likert scale; strongly agree), that would indicate poor RF because optimal RF includes some uncertainty about the content of others' minds. In contrast, Scale B items, uses an ordinal response scale, for instance, an item such as "I'm curious about what others think" where endorsement of response option 6 (strongly agree) indicates optimal RF. However, as outlined below, the alternative scoring schemes have caused a myriad of psychometric challenges for both the adult and adolescent version of the RFQY, leading to abandoning the inclusion of both scoring systems into one measure and instead, focusing efforts on the refinement of subscales independently from each other.

For instance, in Ha, Sharp, Ensink, Fonagy, and Cirino (2013)'s evaluation of the construct validity of the RFQY, the RFQY demonstrated good convergent validity with several measures of mentalizing, including the Child Reflective Function Scale (CRFS; Target et al., 2001) and the Movie for Assessment of Social Cognition (MASC; Dziobek et al., 2006) as well as dimensional and categorical measures of BPD. However, relatively low internal consistency and an unstable factor structure suggested that the RFQY was in need of future psychometric work. Similarly, extensive analyses of both scoring systems using Principal Component Analysis (PCA) and Confirmatory Factor Analysis (CFA) in both community and clinical samples of the adult version of the RFQY failed to yield evidence for the construct validity of either ordinal or median-scored subscales (Fonagy et al., 2016). Fonagy et al. (2016) decided to operationalize the construct of RF in terms of Certainty and Uncertainty and accordingly selected items from Scale A (the median scored items) that lend itself to this conceptualization. Factor analyses resulted in the retention of eight items from the original 26-item Scale A. However, concerns have been raised about the psychometric validity of the RFQ-8 (Müller et al. 2020). First, the same items are scored twice to calculate the respective Certainty and Uncertainty subscale scores, which causes problems for sound psychometric evaluation of the measure in terms of evaluating reliability and the factor structure. Second, the eight items retained in the Fonagy et al. (2016) version refer mostly to understanding one's own mental states and behaviors, while the original definition of mentalizing also includes thinking about others' minds and actions.

In summary, it is clear that the RFQ in general, and the RFQY in particular is in need of further psychometric work. While we support the item reduction methods used in the Fonagy et al., (2016) paper on adults, an alternative to item reduction and measure refinement is through the use of item-level analysis using item response theory (IRT) methods. Thus far, IRT has not been used on either the adult or youth versions of the RFQ. Our goal was to evaluate item performance of ordinal scored RFQ Scale B items given the psychometric challenges posed by using the median-scored items as described above. Our goal was to identify and retain items with satisfactory item characteristics, which also may lead to improved reliability for the measure

(Steinberg & Thissen, 1996). A reduced item set with intuitive scoring would increase the instrument's clinical utility by alleviating time burden in administration and scoring of the measure. To this end, we conducted two studies. Study 1 was conducted in a large sample of 18-25 year old undergraduates. We felt justified using a convenience sample for the initial IRT evaluation because consistent with the idea of prolongation of adolescence, recent work has expanded the definition and timeframe of adolescence to include young adulthood, often up to about 25 years of age (Casey et al., 2010). Moreover, the item adaptation of the adult version of the RFQ was minimal making the RFQY suitable for use in late adolescent and young adult samples.

Study 2 aimed to replicate the derived item pool from Study 1 in a clinical sample of 12-17 year old adolescents. In anticipation of the shortening of the RFQY, Study 2 also aimed to evaluate the construct validity of the new, shorter RFQY by examining its performance against the original 23-item RFQY Scale B in a subsample of clinical adolescents. We examined correlations between scores obtained from the long and shorter versions of the RFQY Scale B, and scores from other measured constructs known to relate to RF in adults and adolescents, including two experimental measures of mentalizing: the MASC (Dziobek et al., 2006) and the Child Eyes Test (Baron Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), BPD features, and emotion dysregulation. In addition, we evaluated whether mean differences would emerge in the expected shortened version of the RFQY between clinical adolescents with high levels of borderline features and healthy controls by comparing RFQY-short scores between the two groups. In summary, our overall aim was to develop a short, effective, intuitively scored, and easy-to-administer version of the RFQY for use in clinical and research settings. Below, we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Study 1

Method

Undergraduate Participants. Data were collected from 737 undergraduate students at a large and racially and ethnically diverse university. Participants were recruited via a mass email advertising an online study to undergraduate students enrolled in at least one Psychology course. The recruitment email was sent from the Department of Psychology and participants self-selected to participate in this study by following a hyperlink to the University's online survey system. Inclusion criteria were English fluency and age between 18 and 25. Participants were informed of the inclusion criteria in a cover letter and were instructed to self-exclude if the aforementioned criteria were not met. The sample included 574 women and 160 men (3 participants did not identify their gender). The mean age in this sample was 21.13 (*SD* = 2.18). The self-identified ethnic breakdown was as follows: Black = 14.8% (*n* = 109), White = 26.5 (*n* = 195), Hispanic = 31.2% (*n* = 230), Asian = 23.2% (*n* = 171), Native American or Alaskan Native = .1% (*n* = 1), and Other = 3.9% (*n* = 29). The majority of the students of this commuter university live at home and work alongside taking classes. This study was approved by the relevant Institutional Review Board and informed consent was obtained. Participants completed questionnaires via a web-based program and were compensated with research credit. All included data is available on the Open Science Framework (https://osf.io/e25p9/?view_only=d534806a666e4583b6649dea662d8d2b).

Measures

The Reflective Function Questionnaire for Youths (RFQY; Sharp et al., 2009). The RFQY is a 46-item self-report measure assessing adolescent reflective function. Participants are asked to rate their responses on a 6-point Likert scale ranging from “Strongly Disagree” to “Strongly Agree”. This measure was adapted from the adult RFQ (Fonagy & Ghinai, unpublished manuscript) which was developed in the United Kingdom and recently validated in adults (Fonagy et al., 2016). Wording of items were rephrased to better fit the U.S. English language and slightly simplified for suitability in youth. The 46-item RFQY demonstrated adequate preliminary psychometric properties in a prior study (Ha, et al., 2013), with good criterion validity with an interview-based measure of reflective function, as well as good convergent and construct validity. Two scoring methods are typically used for items on the original RFQ where one scale (Scale B) consists of 23 Likert-type items with higher numbers indicating higher RF. The other scale (Scale A) consists of scoring based on a median scale, with optimal RF scored at the midpoint of the scale. As mentioned in the introduction, we focused our efforts on the linearly scored Scale B. Therefore, the following 23 items of Scale B were included in the analyses: 7, 13, 14, 15, 23, 26, 32, 38, 3, 4, 6, 11, 18, 19, 20, 21, 34, 39, 41, 42, 43, 44, 45.

Results

The methods of IRT were used to evaluate the 23 items comprising Scale B of the RFQY. The IRT model fitting and the computation of the test statistics were performed using IRTPRO (Version 3) (Cai, du Toit, & Thissen, 2011; Thissen, 2009). Goodness of fit of the IRT models was evaluated using the M_2 statistics and its associated RMSEA value (Maydeu-Olivares & Joe, 2006), and the standardized LD (local dependence) chi-square statistics (based on the LD statistic proposed by (Chen & Thissen, 1997). The graded response model (Samejima, 1969;

1997) was selected as the item response model for these analyses. The sample size in the current study is typical for the IRT model used.

The first IRT analysis included all 23 items using a unidimensional model. In the tables presenting the graded model item parameters, a's refer to slope parameter estimates; b's refer to threshold parameter estimates for a response in a "category or higher." Table 1 presents the graded model item parameter estimates for all 23 items. This analysis showed that ten of the items had little relation (i.e., low slope parameter estimates, less than 1.0) to the underlying construct as measured by the item set, with 80 pairs of items showing local dependence (LD). Local dependence was identified with standardized LD chi-square statistics (a generalization of Chen & Thissen ,1997, as implemented in the IRTPRO software); a standardized chi-square value of 10 or higher was used to "flag" the presence of LD. The presence of LD implies that there is more covariation among the item pairs than can be accounted for by the unidimensional model. To further investigate, a bifactor model that included a 23-item general factor and one specific factor made of the ten items that had little relation to the underlying construct as measured by the 23-item set was used to consider the possibility that those items might form a separate factor. The slope parameter estimates for the specific factor showed substantial loadings for all eight of the reverse-scored items; the remaining two directly-scored items had negative and low slope parameter estimates. The content of the reverse-scored items does not appear to be assessing a construct on which to measure individual differences. For example, some of the items inquire about confusion and difficulty in understanding others' feelings and point of view, while other items seem ambiguous (e.g., "I frequently feel that my mind is empty."). These findings led us to consider the 13 items that showed substantial slope parameter estimates in the 23-item analysis for developing a shorter version of the RFQY.

Table 1

The results of this 13-item analysis (Table 2) showed that while all the items had substantial slope parameter estimates, implying strong relation with the underlying construct, there were 12 pairs of items showed LD. With only 13 items, decisions for item retention were mainly based on item content with some consideration given to the magnitude of the slope parameter and presence of LD. We opted to omit the following items: (1) Item 4, “I realize that I can sometimes misunderstand my best friends’ reactions” because the “sometimes” qualification may draw endorsement regardless of RF, (2) Item 6, “Other people tell me I’m a good listener” requires “other people” to communicate with respondent, (3) Item 20, “Understanding the reasons for people’s actions helps me to forgive them” includes the idea of forgiving others, (4) Items 39, 42, 43, and 44 have content that seems tangential to the core construct of RF and three item-pairs that include items 39, 42, and 44 show LD. (These items in order are: “In order to know exactly how someone is feeling, I have found that I need to ask them,” “I have noticed that people often give advice to others that they actually wish to follow themselves,” “I wonder what my dreams mean,” and “How I feel can easily affect how I understand someone else’s behavior.”)

Table 2

The remaining six items were analyzed separately; item parameters are listed in Table 3. The unidimensional model had adequate fit, (M_2 (369) = 2270.33, p < .0001; RMSEA = .08). However, one item-pair showed evidence of LD (“I pay attention to my feelings.” and “In an argument, I keep the other person’s point of view in mind.”). There are two strategies for dealing with LD. One strategy is to remove one of the items of the pair that show LD from the analysis. Another strategy is to evaluate the significance of the LD and if significant, combine the responses to the two items into a testlet. The significance of LD was evaluated using a bifactor model with an equal slope specific factor. The equal-slope specific factor provided significant improvement in model fit (G^2 (1) = 23.99, $p < .001$); the residual correlation is .15. Because these two items provide a better balance of content, we opted to retain the items in the short version of the RFQY. A testlet is made of the sum of the item responses to the two items, thus, creating a single "super" item (Steinberg & Thissen, 1996; Thissen & Steinberg, 2010). The testlet is used for item parameter estimation so that the slope parameters are not influenced by the excess covariation between items 18 and 19. Table 4 presents the item parameters for the four items and the testlet. The use of the testlet does not affect the traditional sum score, as it is made up of the sum of the item responses.

Table 3

Table 4

Study 2

The IRT analysis of the RFQY identified six items for a shorter version of the RFQY. In Study 2, we aimed to replicate the IRT analyses and provide evidence of construct validity for the shortened version.

Method

Participants. This study included a sample of 12–17-year-olds admitted to the adolescent unit of a private psychiatric hospital and a sample of high school students. Adolescent patients were eligible for study inclusion if they were admitted to the adolescent unit and had sufficient fluency in English to complete all research assessments. Patients who were ineligible for study participation had a diagnosis of schizophrenia or any psychotic disorder, and/or met criteria for an intellectual disability. Parents provided informed consent and adolescents provided informed assent. The dataset included a total of 567 consecutively admitted adolescents (66.4% female;

mean age 15.30; SD = 1.47). We randomly selected 100 subjects from the 567 that were reserved for the analyses directed to establish evidence of construct validity. These 100 subjects were therefore not included in the IRT replication and the final IRT sample consisted of 467 adolescents.

Healthy controls were recruited through high schools and the general community. A total of N = 186 adolescents participated in the study. However, there were missing data on key study measures that reduced the sample size to 128 participants (60.1% female; mean age 15.42; SD = 1.25). All included data is available on the Open Science Framework (https://osf.io/e25p9/?view_only=d534806a666e4583b6649dea662d8d2b).

Measures

The Reflective Function Questionnaire for Youths (RFQY; Sharp et al., 2009). We used the same 23 items forming Scale B of the instrument described in Study 1.

The Movie Assessment for Social Cognition (MASC; Dziobek et al., 2006)**.** The MASC is a computerized test for the assessment of implicit theory of mind or mentalizing abilities that approximates the demands of everyday life (Smeets, Dziobek, & Wolf, 2009). Participants are asked to watch a 15-minute film about four characters getting together for a dinner party. Themes of each segment cover friendship and dating issues. During administration of the task, the film is stopped at 45 points during the plot and questions referring to the characters' mental states (feelings, thoughts, and intentions) are asked (e.g., “What is Betty feeling?”, “What is Cliff thinking?”). While subscales scores are available, a total score (number correct) was employed in the current study. The MASC is a reliable instrument that has proven sensitive in detecting subtle mindreading difficulties in adults of and adolescents (Dziobek et al., 2006) (Smeets et al., 2009).

The Children's Eyes Test. Child's Eye Task (CET; Baron Cohen et al., 2001), includes 28 black and white photographs of the eye region and the participant is asked to pick one of four words that best describes what the person in the photo is thinking or feeling. Three of the four words are foil mental state terms, while the fourth is deemed ‘‘correct.’’ The position of the correct answer is randomized for each item. Correct answers are scored as 1 and then summed to produce a total "correct" score. In contrast to the MASC, the CET is considered a single mode (emotion recognition) and explicit measure of mentalizing because the participant is instructed to read the mind of an individual through explicitly focusing on the eyes (Sharp et al., 2013).

Borderline Personality Disorder Features Scale for Children. The Borderline Personality Features Scale for Children (BPFSC; Crick, Murray-Close, & Woods, 2005) is a 24-item questionnaire measure that assesses borderline personality features in children ages nine and older, including adolescents. The measure was adapted from the borderline subscale of the Personality Assessment Inventory (PAI; Morey, 1991)), which has been shown to be a valid and reliable measure. Responses are scored on a 5-point Likert-type scale, ranging from 1 (not at all true) to 5 (always true) with higher total scores indicating greater levels of borderline personality features. The BPFSC has been identified as a useful tool in assessing borderline pathology in adolescents (Chang, Sharp, & Ha, 2011). The internal consistency of the BPFSC in the current study was high (α = .91).

The Difficulties in Emotion Regulation Scale. The Difficulties in Emotion Regulation Scale (DERS; (Gratz & Roemer, 2004)) is a self-report questionnaire measure that assesses emotion dysregulation. It consists of 36 items that are scored on a 5-point Likert scale, ranging from 1 (*'almost never'*) to 5 (*'almost always)'*). A higher score indicates greater emotion dysregulation. The measure assesses six separate scales including: *nonacceptance of emotions,* *difficulties in goal-directed behavior, impulse control difficulties, lack of emotional awareness, limited access to strategies, and lack of emotional clarity*. For the current study, we used the subscale scores of the DERS given that the scale was originally developed to measure six components of difficulties in emotion regulation. In the measure's initial publication, the DERS displayed good internal consistency (α = .93), construct and predictive validity, and test-retest reliability across 4 to 8 weeks ($p < .01$) (Gratz & Roemer, 2004) and has been validated for use in adolescents (Perez, Venta, Garnaat, & Sharp, 2012). In the present study, the measure had good internal consistency with Cronbach's alpha for each of the subscales as follows: nonacceptance of emotions (α = .91), difficulties in goal-directed behavior (α = .91), impulse control difficulties (α = .93), lack of emotional awareness (α = .85), limited access to strategies (α = .93), and lack of emotional clarity (α = .87).

Results

Replication in clinical sample. IRT analyses reported as part of Study 1 identified six items to comprise the brief version of the RFQY. We investigated whether these items performed comparably with item responses from the adolescent clinical sample. The six-item unidimensional analysis showed adequate fit (M_2 (369) = 887.14, p < .0001; RMSEA = .05), with no evidence of local dependence. However, as shown in Table 5, Item 11 has a low slope parameter indicating that the item is not strongly related to the underlying construct as defined by the other 5 items. Possibly, the wording of the item, including the phrases "I believe" and "based on their own beliefs" is cognitively challenging for the younger clinical sample, thus the item might be open to misinterpretation. Consequently, the shortened version of the RFQY would include the five items that show substantial slopes in both the undergraduate and clinical samples.

Construct validity. Next, we investigated whether scores obtained on the RFQY-5 show similar patterns of relations with other variables that are observed with scores based on the original 23-item set.

Table 5

First, in terms of reliability, we note that the Cronbach's alpha in the sub-sample of 100 clinical adolescents was .75 for the RFQY-5. For the 23-item set it was .69. A reduced item set show can show greater measurement precision than the original longer version, particularly when items that add little information are removed (Steinberg & Thissen;1996). Next, we examined correlations between scores obtained from the original and short versions of the RFQY, and scores from other measured constructs known to relate to RF in adults and adolescents, including two experimental measures of mentalizing: the MASC (Dziobek et al., 2006) and the Child Eyes Test (Baron Cohen et al., 2001), BPD features, and emotion dysregulation. In addition, we evaluated whether the shortened RFQY would discriminate between adolescents with high vs. low borderline features by comparing RFQY-5 scores between adolescent inpatients and healthy adolescent controls. Table 6 contains the results of correlational analyses and shows similar patterns of correlations for the RFQY-5 as the original 23-item Scale B in relation to experimental measures of mentalizing (the MASC and the CET). Using Lee and Preacher's (2013) implementation of Steiger's (1980) equations to evaluate the significance of the difference between two dependent correlations, no differences in correlations were detected among the RFQY-5 and the original Scale B relations with MASC and CET, respectively. As recommended by Counsell and Cribbie (2015), we used Anderson and Hauck's (1983) formula to evaluate equivalence between two dependent correlations. For all equivalence tests, the smallest effect size of interest (SESOI) was set to r = 0.191 based on power analyses conducted by Lakens (2017) for 80% power in a sample of 100 participants with an alpha of .05. The equivalence test for the correlations was significant for both the MASC (*p* < .01) and the CET (*p* = .02).

Correlations with the DERS revealed similar discriminant validity for the RFQY-5 compared to other scales. Correlations of RFQY-5 and the original Scale B with the subscales of the DERS are nonsignificant for Non-Acceptance of Emotions, Difficulties in Goal-Directed Behavior, and Limited Access to Strategies subscales. The RFQ-5 had a significant negative (although small in magnitude) correlation with Impulse control difficulties, while the original Scale B did not. Other correlations with DERS subscales are more substantial; for example, the correlations of scores obtained from the RFQY-5 and Scale B with Lack of Emotional Awareness of -.62 and -.58 are relatively strong. Applying a test for significance using Lee and Preacher's software revealed differences between the correlation between RFQY-5 and the original Scale B with Non-Acceptance of Emotions (*z* = -1.973, *p* = .048) and with Impulse Control Difficulties (*z* = -1.996, *p* = .046), but no other DERS subscale. Similarly, the Anderson Hauck equivalence test was significant for all subscales (all *p*s < .05) except the Non-Acceptance of Emotions (*p* = .146) and Impulse Control Difficulties (*p* = .130) subscales.

Correlations between the RFQ-5 and both parent-and self-reported borderline features were of similar magnitude as the original Scale B scores. Using Lee and Preacher's software, the difference between the correlations of the RFQY-5 (r = -.30) and Scale B (r = -.25) with self-report BPD features was not significant (*z* = 0.786, *p* = .43); the correlations of RFQY-5 and Scale B with parent-report BPD features (r = -.28 and r = -.25, respectively) are also not significantly different (*z* = -0.584, *p* = .55). The Anderson Hauck equivalence test was significant for both parent (*p* < .01) and self (*p* = .01) reports.

Table 6

To examine mean differences between inpatient adolescents (n = 100) and healthy controls (n = 183), independent samples t-tests were conducted using the BPFSC, RFQY-5, and the original Scale B. These results are displayed in Table 7. There was a significant difference on the BPFSC between the two groups; clinical adolescents had significantly higher levels BPD features than healthy control adolescents. The RFQY-5, but not the 23-item Scale B discrimniated distinguished between the two groups.

Table 7

Discussion

The aim of the current study was to improve the validity and reliability of the RFQY Scale B by first conducting IRT analyses in a non-clinical sample of young people to identify poor functioning items, followed by a replication study in a clinical sample in addition to an investigation of evidence for construct validity to evaluate whether a shortened scale performed as effectively as longer, more time-consuming version of the scale. Against the background of increased studies of reflective function in general (see Katznelson, 2014 for a review), and the use of the RFQ in both adults and adolescents (e.g. Badoud et al., 2015; Badoud et al., under review; Bo et al., 2016; Fonagy et al., 2016; Ha et al., 2013; Somma, Borroni, Drislane, & Fossati, 2016), further validity work is warranted especially because the current study is only the second study to evaluate the psychometric properties of the RFQY. In addition, reducing the number of items of the RFQY would improve the instrument's clinical utility by alleviating time burden in both administration and scoring of the measure.

The IRT item analyses of the 23 Scale B items showed that many of the items had little relation to the underlying construct and 80 item pairs had LD that indicated redundancy of item content. A subset of 13 items that showed sufficient relation to the underlying construct was then evaluated for inclusion in a shorter version of the RFQY. This analysis also revealed indices of LD and items were removed based mostly on considerations of item content. A final set of six items formed the short version. One (Item 11) of the six items did not perform well in the analysis of the item response data from the younger clinical sample. As a consequence, the brief measure includes five items that had substantial slope parameter estimates in both samples.

Analyses aimed at evaluating the construct validity of the short version compared with the original 23-item version (Scale B), suggest that the RFQY-5 performs equally to the long form in terms of correlations with other variables, and outperformed the long version in terms of discriminating between inpatient adolescents with high levels of borderline traits compared to community-dwelling adolescents with significantly lower levels of borderline traits.

It is worth considering the implications of the current findings for understanding the construct meaning of RF as it is reflected in the five items retained through the IRT analyses. As mentioned in the introduction, RF is a multicomponent concept that includes dimensions of cognition (thoughts) and affect (feelings), self and other, as well as implicit and explicit mentalizing. The items retained in the RFQY-5 (I believe that people can see a situation very differently based on their own beliefs and experiences; I pay attention to my feelings; In an argument, I keep other person’s points of view in mind; I like to think about reasons behind my actions; I’m often curious about the meaning behind others’ actions; and I pay attention to the impact of my actions on others’ feelings) appear to have retained coverage of content that relate to cognition rather than affect. Affect is covered in two items, but in the context of “thinking about feelings” (I pay attention to my feelings; I pay attention to the impact of my actions on others’ feelings). In this sense, the items retained in the RFQY-5 closely resemble the original operationalization of RF in Fonagy's (1991) early descriptions of BPD as a mentalizing disorder where RF is described as a metacognitive capacity ("thinking about thinking"). This operationalization is also very close to the original definition of second-order theory of mind (Happé, 1994).

Our results furthermore show that the five items retained in the RFQY-5 appear to cover thinking about both self and other, but appear to lose coverage of implicit mentalizing. Put differently, most items retained in the RFQY-5 relate to explicit mentalizing that lies within conscious awareness. We acknowledge, however, that self-report measures, by their very nature, would be biased towards content of explicit mentalizing and that implicit mentalizing is probably best captured by experimental tasks (Sharp et al., 2013).

The five items identified through the current analyses do not overlap with the items identified in the recent validation paper of the RFQ in adults (Fonagy et al., 2016), because that scale was developed using Scale A items. Against the background of psychometric difficulties associated with the 46 item measure, Fonagy et al. (2016) recoded Scale A RFQ items to either reflect *Certainty* in mental states of self and other (e.g. "I always know what I feel" or *Uncertainty* in the mental states of self and other ("Sometimes I do things without really knowing why"). Certainty, in this context, was described as "hypermentalizing" elsewhere defined as overattribution of mental states to self/others in the absence of evidence to support conclusions (Sharp, 2014; Sharp & Vanwoerden, 2015) while Uncertainty was described as "hypomentalizing" (defined as a deficiency in mentalizing capacity). Fonagy and colleagues retained eight items in their analyses (People's thoughts are a mystery to me**;** I don't always know why I do what I do; When I get angry I say things without really knowing why I am saying them; When I get angry I say things that I later regret; If I feel insecure I can behave in ways that put others’ backs up; Sometimes I do things without really knowing why; I always know what I feel; Strong feelings often cloud my thinking). In evaluating and comparing this item set with the one derived in the current study for the RFQY-5, it appears that the 8-item adult RFQ is more affectively loaded with stronger coverage of the emotion dysregulation consequences of poor RF. It is concerned more strongly with self-regulatory processes associated with mentalizing than the RFQY-5 item content. Fonagy et al. (2016) did not make use of IRT to reduce the item set, but rather took a more content-driven approach to refining the questionnaire. To further advance the psychometric development and refinement of the RFQ, both in adults, young adults and adolescents, it is important that future studies compare and contrast the eight-item measure derived through the content driven approach used by Fonagy et al. (2016), with the five-item version suggested through the IRT methods employed here.

The study has several limitations. The IRT analyses were conducted using item response data obtained from a sample of undergraduate college students which is consistent with the notion of the prolongation of adolescence. However, future work would benefit from research in unselected samples of adolescents in the community. In addition, our clinical sample was drawn from an inpatient setting that limits generalizability to other clinical populations. And while the college sample and community-based sample of adolescents were highly diverse, the clinical sample was not. Therefore, future research should include more diverse clinical samples from multiple treatment settings. In addition, gender was not well represented in the undergraduate sample, which is often the case in undergraduate samples of psychology majors. Gender was, however, well represented in the adolescent clinical and community-based samples. Finally, while the focus in the current study was on refinement of Criterion B given its advantages discussed earlier, unfolding or ideal-point IRT models used of mid-point scoring may be appropriate to evaluate Scale A in future work.

Despite these limitations, the current study is important in that it produced a shortened, streamlined version of the RFQY which appears to capture the essence of the RF construct as originally intended, and which provides high measurement precision due to careful IRT analyses that eliminated poor functioning items. The improved measurement precision, combined with the ease of administration of a short self-report measure, significantly increases the clinical utility of reflective function which is gaining more momentum as a malleable target of treatment that cuts across most therapeutic modalities (Fonagy & Luyten, 2016; Katznelson, 2014).

References

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. Communications in Statistics-Theory and Methods, 12(23), 2663-2692.

Badoud, D., Luyten, P., Fonseca-Pedrero, E., Eliez, S., Fonagy, P., & Debbane, M. (2015). The French Version of the Reflective Functioning Questionnaire: Validity Data for Adolescents and Adults and Its Association with Non-Suicidal Self-Injury. *Plos One, 10*(12). doi: ARTN e0145892 10.1371/journal.pone.0145892

Badoud, D., Perroud, N., Prada, P., Nicastro, R., Gremond, C., & Luyten, P. (under review). Attachment and reflective functioning in women with borderline personality disorder.

Baron Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the mind in the eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 42*(2), 241-251.

Bateman, A. W., & Fonagy, P. (2004). Mentalization-based treatment of BPD. *Journal Personality Disorders, 18*(1), 36-51.

Bo, S., Sharp, C., Beck, E., Pedersen, J., Gondan, M., & Simonsen, E. (2016). First Empirical Evaluation of Outcomes for Mentalization-Based Group Therapy for Adolescents With BPD. *Personality Disorders: Theory, Research, and Treatment*.

Bouchard, M. A., Target, M., Lecours, S., Fonagy, P., Tremblay, L. M., Schachter, A., & Stein, H. (2008). Mentalization in adult attachment narratives: Reflective functioning, mental states, and affect elaboration compared. *Psychoanalytic Psychology, 25*(1), 47-66. doi: 10.1037/0736-9735.25.1.47

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: Flexible professional item response theory modeling for patient reported outcomes [Computer software]*. Chicago: SSI International.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2(P) tables. *British Journal of Mathematical & Statistical Psychology, 59*, 173-194. doi: 10.1348/000711005X66419

Casey BJ, Duhoux S, Malter Cohen M. Adolescence: What do transmission, transition, and translation have to do with it? Neuron. 2010;67:749–60.Chanen, A. M. (2015). Borderline Personality Disorder in Young People: Are We There Yet? *J Clin Psychol, 71*(8), 778-791. doi: 10.1002/jclp.22205

Chanen, A. M., & Kaess, M. (2012). Developmental pathways to borderline personality disorder. [Research Support, Non-U.S. Gov't Review]. *Curr Psychiatry Rep, 14*(1), 45-53. doi: 10.1007/s11920-011-0242-y

Chang, B., Sharp, C., & Ha, C. (2011). The criterion validity of the Borderline Personality Feature Scale for Children in an adolescent inpatient setting. *Journal of Personality Disorders, 25*(4), 492-503.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs: Using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289. doi: Doi 10.3102/10769986022003265

Counsell, A., & Cribbie, R. A. (2015). Equivalence tests for comparing correlation and regression coefficients. British Journal of Mathematical and Statistical Psychology, 68(2), 292-309.

Crick, N. R., Murray-Close, D., & Woods, K. (2005). Borderline personality features in childhood: a short-term longitudinal study. *Development and Psychopathology, 17*(4), 1051-1070.

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., . . . Convit, A. (2006). Introducing MASC: a movie for the assessment of social cognition. *J Autism Dev Disord, 36*(5), 623-636. doi: 10.1007/s10803-006-0107-0

Fonagy, P. (1991). Thinking about thinking: Some clinical and theoretical considerations in the treatment of a borderline patient. *International Journal of Psycho-Analysis, 72*, 639-656.

Fonagy, P., Gergely, G., Jurist, E. L., & Target, M. (2002). *Affect regulation, mentalization, and the development of self*. New York: Other Press.

Fonagy, P., & Ghinai, R. (Unpublished manuscript). *A self-report measure of mentalizing: Development and preliminary test of the reliability and validity of the Reflective Function Questionnaire (RFQ)*. University College London.

Fonagy, P., & Luyten, P. (2016). A multilevel perspective on the development of borderline personality disorder. In D. C. Cichetti (Ed.), *Developmental psychopathology* (Vol. 3, pp. 726-792). Hoboken, New Jersey: John Wiley and Sons.

Fonagy, P., Luyten, P., Moulton-Perkins, A., Lee, Y. W., Warren, F., Howard, S., . . . Lowyck, B. (2016). Development and Validation of a Self-Report Measure of Mentalizing: The Reflective Functioning Questionnaire. *Plos One, 11*(7), e0158678. doi: 10.1371/journal.pone.0158678

Fonagy, P., Speranza, M., Luyten, P., Kaess, M., Hessels, C., & Bohus, M. (2015). ESCAP Expert Article: borderline personality disorder in adolescence: an expert research review with implications for clinical practice. [Research Support, Non-U.S. Gov't Review]. *Eur Child Adolesc Psychiatry, 24*(11), 1307-1320. doi: 10.1007/s00787-015-0751-z

Fonagy, P., Steele, H., Moran, G., Steele, M., & Higgitt, A. (1991). The capacity for understanding mental states: The reflective self in parent and child and its significance for securtiy of attachment. *Infant Mental Health Journal, 12*(3), 201-218.

Fonagy, P., & Target, M. (1997). Attachment and reflective function: Their role in self-organization. *Development and Psychopathology, 9*(4), 697-700.

George, C., Kaplan, N., & Main, M. (1985). *The Berkeley Adult Attachment Interview*: Unpublished protocol, Department of Psychology, University of California, Berkeley.

Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure and initial validation of the Difficulties in Emotion Regulation Scale. *Journal of Psychopathology and Behavioral Assessment, 26*(1), 41-54.

Ha, C., Sharp, C., Ensink, K., Fonagy, P., & Cirino, P. (2013). The measurement of reflective function in adolescents with and without borderline traits. *J Adolesc, 36*(6), 1215-1223. doi: DOI 10.1016/j.adolescence.2013.09.008

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154.

Hill, L. L., Levy, K. N., Meehan, K. B., & Reynoso, J. S. (2007). Reliability of a multidimensional measure for scoring reflective function. [Validation Studies]. *J Am Psychoanal Assoc, 55*(1), 309-313.

Katznelson, H. (2014). Reflective functioning: a review. [Review]. *Clin Psychol Rev, 34*(2), 107-117. doi: 10.1016/j.cpr.2013.12.003

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Social psychological and personality science, 8(4), 355-362.

Lee, I. A., & Preacher, K. J. (2013, October). Calculation for the test of the difference between two dependent correlations with no variable in common [Computer software]. Available from http://quantpsy.org.

Levy, K. N., Meehan, K. B., Kelly, K. M., Reynoso, J. S., Weber, M., Clarkin, J. F., & Kernberg, O. F. (2006). Change in attachment patterns and reflective function in a randomized control trial of transference-focused psychotherapy for borderline personality disorder. *Journal of Consulting and Clinical Psychology*, *74*(6), 1027-1040. doi: 10.1037/0022-006X.74.6.1027

Maydeu-Olivares, A. & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, 71, 713-732. DOI: 10.1007/s11336-005-1295-9.

Meehan, K. B., Levy, K. N., Reynoso, J. S., Hill, L. L., & Clarkin, J. F. (2009). Measuring reflective function with a multidimensional rating scale: comparison with scoring reflective function on the AAI. [Comparative Study]. *J Am Psychoanal Assoc, 57*(1), 208-213. doi: 10.1177/0003065109057001108

Morey, L. (1991). *Personality Assessment Inventory*. Odessa, FL: Psychological Assessment Resources

Pajulo, M., Pyykkonen, N., Kalland, M., Sinkkonen, J., Helenius, H., Punamaki, R. L., & Suchman, N. (2012). Substance-abusing mothers in residential treatment with their babies: Importance of pre- and postnatal maternal reflective functioning. *Infant Mental Health Journal, 33*(1), 70-81. doi: 10.1002/imhj.20342

Perez, J., Venta, A., Garnaat, S., & Sharp, C. (2012). The Difficulties in Emotion Regulation Scale: Factor Structure and Association with Nonsuicidal Self-Injury in Adolescent Inpatients. *Journal of Psychopathology and Behavioral Assessment, 34*(3), 393-404. doi: DOI 10.1007/s10862-012-9292-7

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika, 34*(4p2), 1-&.

Samejima, F. (1997). Graded response model. In W. van der Linden & J. R. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85-100). New York: Springer.

Schechter, D. S., Myers, M. M., Brunelli, S. A., Coates, S. W., Zeanah, C. H., Davies, M., . . . Liebowitz, M. R. (2006). Traumatized mothers can change their minds about their toddlers: Understanding how a novel use of videofeedback supports positive change of maternal attributions. *Infant Mental Health Journal, 27*(5), 429-447. doi: 10.1002/imhj.20101

Sharp, C. (2014). The social-cognitive basis of borderline personality disorder: A theory of hypermentalizing. In C. Sharp & J. Tackett (Eds.), *The handbook of borderline personality disorder in children and adolescents*. New York: Springer.

Sharp, C. (2016). Bridging the gap: the assessment and treatment of adolescent personality disorder in routine clinical care. [Review]. *Arch Dis Child*. doi: 10.1136/archdischild-2015-310072

Sharp, C., & Fonagy, P. (2015). Practitioner review: Emergent borderline personality disorder in adolescence: – Recent conceptualization, intervention, and implications for clinical practice. *Journal of Child Psychology and Psychiatry*(56(12)), 1266-1288.

Sharp, C., Ha, C., Carbone, C., Kim, S., Perry, K., Williams, L., & Fonagy, P. (2013). Hypermentalizing in adolescent inpatients: treatment effects and association with borderline traits. *J Pers Disord, 27*(1), 3-18. doi: 10.1521/pedi.2013.27.1.3

Sharp, C., Ha, C., Michonski, J., Venta, A., & Carbone, C. (2012). Borderline personality disorder in adolescents: evidence in support of the Childhood Interview for DSM-IV Borderline Personality Disorder in a sample of adolescent inpatients. *Comprehensive Psychiatry, 53*(6), 765-774. doi: 10.1016/j.comppsych.2011.12.003

Sharp, C., & Romero, C. (2007). Borderline personality disorder: a comparison between children and adults. *Bull Menninger Clin, 71*(2), 85-114. doi: 10.1521/bumc.2007.71.2.85

Sharp, C., & Vanwoerden, S. (2015). Hypermentalizing in Borderline Personality Disorder: A model and data. *Journal of Infant, Child, and Adolescent Psychotherapy, 14*(1), 33-45.

Sharp, C., Williams, L. W. W., Ha, C., Baumgardner, J., Michonski, J., Seals, R., . . . Fonagy, P. (2009). The development of a mentalization-based outcomes and research protocol for an adolescent in-patient unit. *The Bulletin of the Menninger Clinic*.

Shmueli-Goetz, Y., Target, M., Fonagy, P., & Datta, A. (2008). The Child Attachment Interview: a psychometric study of reliability and discriminant validity. *Dev Psychol, 44*(4), 939-956. doi: 2008-08592-005 [pii] 10.1037/0012-1649.44.4.939

Smeets, T., Dziobek, I., & Wolf, O. T. (2009). Social cognition under stress: Differential effects of stress-induced cortisol elevations in healthy young men and women. *Horm Behav*. doi: S0018-506X(09)00026-9 [pii] 10.1016/j.yhbeh.2009.01.011

Somma, A., Borroni, S., Drislane, L. E., & Fossati, A. (2016). Assessing the Triarchic Model of Psychopathy in Adolescence: Reliability and Validity of the Triarchic Psychopathy Measure (TriPM) in Three Samples of Italian Community-Dwelling Adolescents. *Psychological Assessment, 28*(4), E36-E48. doi: 10.1037/pas0000184

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*(1), 81-97. doi: Doi 10.1037/1082-989x.1.1.81

Target, M., Fonagy, P., Shmueli-Goetz, Y., Data, A., & Schneider, T. (2007). *The Child Attachment Interview (CAI) Protocol*. London: University College London.

Target, M., Oandasan, C., & Ensink, K. (2001). *Child Reflective Functioning Scale scoring manual: For application to the child attachment interview.* . Anna Freud Centre/University College London. : Unpublished manuscript.

Thissen, D. (2009). *The MEDPRO project: An SBIR project for a comprehensive IRT and CAT software system-IRT software.* Paper presented at the Proceedings of the 2009 GMAC conference on computerized adaptive testing.

Thissen, D., & Steinberg, L. (2010). Using item response theory to disentangle constructs at different levels of generality. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 123-144). Washington, DC: American Psychological Association.

Toth, S. L., Rogosch, F., & Cicchetti, D. (2008). Attachment-theory-informed intervention and reflective functioning in depressed mothers. In H. Steele & M. Steele (Eds.), *Clinical applications of the adult attachment interview* (pp. 154-172). New York: Guilford Press.

Zanarini, M. C. (2003). *The Child Interview for DSM-IV Borderline Personality Disorder*. Belmont, MA: McLean Hospital.

Table 1. Undergraduate Sample, Graded Model Item Parameter Estimates for the 23 RFQY Scale B items

Item	a	s.e.	b1	s.e.	b2	s.e.	b3	s.e.	b4	s.e.	b5	s.e.
3	0.75	0.08	-2.78	0.29	-1.49	0.18	-0.38	0.11	1.50	0.19	3.76	0.41
4	0.99	0.09	-2.80	0.24	-1.71	0.15	-0.75	0.10	1.17	0.13	3.45	0.32
6	1.83	0.12	-2.61	0.17	-1.99	0.12	-1.11	0.08	-0.08	0.06	0.86	0.08
7(R)	0.29	0.07	-14.01	3.56	-9.19	2.29	-3.51	0.90	-0.32	0.27	3.93	0.99
11	2.33	0.15	-2.08	0.12	-1.73	0.10	-1.03	0.07	-0.19	0.05	0.60	0.06
13(R)	0.53	0.08	-8.49	1.32	-5.54	0.80	-2.28	0.34	-0.31	0.15	1.91	0.30
20	2.44	0.15	-2.32	0.13	-1.85	0.10	-0.98	0.06	0.07	0.05	1.03	0.07
15(R)	0.43	0.07	-8.24	1.46	-6.11	1.06	-2.54	0.46	-0.18	0.18	2.66	0.48
18	1.80	0.12	-3.05	0.21	-2.03	0.12	-1.07	0.08	0.03	0.06	1.25	0.09
21	0.46	0.07	-4.67	0.75	-2.47	0.41	-0.08	0.17	2.47	0.42	5.68	0.92
14(R)	0.48	0.08	-8.24	1.35	-5.41	0.85	-2.12	0.36	-0.29	0.16	2.05	0.35
19	1.74	0.11	-2.73	0.18	-1.97	0.12	-0.98	0.07	0.27	0.06	1.52	0.11
23(R)	0.12	0.07	-28.66	16.91	-17.73	10.44	-5.62	3.35	1.35	1.00	10.17	6.00
26(R)	0.01	0.07	-471.20	4683.77	-267.90	2662.93	-93.92	933.54	23.87	237.50	183.56	1824.55
32(R)	0.50	0.08	-7.28	1.14	-4.99	0.76	-2.06	0.34	-0.56	0.17	1.35	0.25
34	2.31	0.14	-2.23	0.13	-1.59	0.09	-0.95	0.06	0.16	0.05	1.39	0.09
38(R)	-0.03	0.07	133.51	359.25	69.40	186.73	13.72	37.03	-20.67	55.67	-53.38	143.62
39	1.18	0.09	-2.80	0.21	-1.79	0.14	-0.67	0.08	0.71	0.09	2.28	0.18
41	2.12	0.13	-2.17	0.13	-1.62	0.09	-0.96	0.07	0.16	0.05	1.18	0.08
42	2.32	0.15	-2.24	0.13	-1.62	0.09	-0.91	0.06	0.06	0.05	1.11	0.08
43	1.47	0.10	-2.45	0.17	-1.81	0.12	-1.05	0.09	-0.00	0.06	1.12	0.10
44	1.68	0.11	-2.27	0.15	-1.83	0.12	-0.94	0.08	0.31	0.06	1.62	0.11
45	3.25	0.21	-2.18	0.12	-1.61	0.08	-1.01	0.06	-0.05	0.04	0.86	0.06

Note: The threshold parameters cannot be estimated when the item has near zero slope parameter estimates (see for example item 26); thus, for those items, the estimates for the thresholds are out of range and not interpretable.

Table 2. Undergraduate Sample, Graded Model Item Parameter Estimates for 13 of the RFQY Scale B items

Item	a	s.e.	b1	s.e.	b2	s.e.	b3	s.e.	b4	s.e.	b5	s.e.
4	1.03	0.09	-2.72	0.23	-1.65	0.15	-0.72	0.10	1.14	0.12	3.34	0.30
6	1.78	0.12	-2.66	0.17	-2.01	0.12	-1.11	0.08	-0.06	0.06	0.87	0.08
11	2.31	0.15	-2.10	0.12	-1.73	0.10	-1.01	0.07	-0.17	0.05	0.61	0.06
18	1.75	0.11	-3.14	0.21	-2.06	0.13	-1.07	0.08	0.05	0.06	1.26	0.10
34	2.33	0.15	-2.24	0.13	-1.58	0.09	-0.92	0.06	0.18	0.05	1.37	0.09
19	1.71	0.11	-2.78	0.18	-1.98	0.12	-0.97	0.08	0.28	0.06	1.52	0.11
20	2.39	0.15	-2.35	0.14	-1.85	0.10	-0.96	0.06	0.08	0.05	1.03	0.08
39	1.24	0.09	-2.72	0.21	-1.73	0.13	-0.64	0.08	0.70	0.09	2.20	0.17
41	2.23	0.14	-2.13	0.13	-1.57	0.09	-0.92	0.07	0.16	0.05	1.15	0.08
42	2.43	0.15	-2.22	0.13	-1.59	0.09	-0.87	0.06	0.07	0.05	1.08	0.08
43	1.52	0.11	-2.40	0.16	-1.77	0.12	-1.02	0.08	0.01	0.06	1.09	0.09
44	1.78	0.12	-2.20	0.14	-1.77	0.11	-0.90	0.07	0.31	0.06	1.56	0.11
45	3.29	0.21	-2.20	0.12	-1.60	0.08	-0.98	0.06	-0.03	0.05	0.85	0.06

Table 3. Undergraduate Sample, Graded Model Item Parameter Estimates for 6 of the RFQY Scale B items

Item	a	s.e.	b1	s.e.	b2	s.e.	b3	s.e.	b4	s.e.	b5	s.e.
11	2.21	0.15	-2.13	0.13	-1.77	0.10	-1.05	0.07	-0.19	0.06	0.62	0.06
18	1.88	0.13	-2.97	0.20	-2.00	0.12	-1.05	0.08	0.03	0.06	1.23	0.09
19	1.85	0.12	-2.64	0.17	-1.90	0.12	-0.95	0.07	0.26	0.06	1.48	0.10
34	2.47	0.16	-2.18	0.13	-1.56	0.09	-0.93	0.06	0.16	0.05	1.36	0.08
41	1.87	0.13	-2.30	0.14	-1.71	0.11	-1.01	0.08	0.16	0.06	1.26	0.09
45	3.05	0.22	-2.22	0.12	-1.65	0.09	-1.02	0.06	-0.06	0.05	0.88	0.07

Table 4. Undergraduate Sample, Graded Model Item Parameter Estimates for 4 of the RFQY Scale B items and the testlet of items 18 and 19

Item	a	s.e.	b1	s.e.	b2	s.e.	b3	s.e.	b4	s.e.	b5	s.e.
11	2.21	0.15	-2.13	0.13	-1.77	0.10	-1.05	0.07	-0.19	0.05	0.62	0.06
34	2.47	0.17	-2.19	0.13	-1.57	0.09	-0.93	0.06	0.16	0.05	1.37	0.08
41	1.94	0.13	-2.27	0.14	-1.69	0.10	-1.00	0.07	0.16	0.06	1.24	0.09
45	3.15	0.23	-2.21	0.12	-1.64	0.08	-1.02	0.06	-0.06	0.05	0.88	0.06
testlet	2.07	0.13	-3.06	0.20	-2.81	0.18	-2.23	0.13	-1.89	0.11	-1.32	0.08
18 + 19			b6	s.e.	b7	s.e.	b8	s.e.	b9	s.e.	b10	s.e.
			-0.79	0.06	-0.05	0.06	0.43	0.06	1.20	0.08	1.83	.11

Note: The testlet has 11 response categories, estimating 10 threshold parameters for a category or higher; these parameter estimates are presented in two lines. The items retained for the RFQY-Short are: 11. I believe that people can see a situation very differently based on their own beliefs and experiences, 18. I pay attention to my feelings, 19. In an argument, I keep other person's points of view in mind, 34. I like to think about reasons behind my actions, 41. I'm often curious about the meaning behind others' actions, and 45. I pay attention to the impact of my actions on others' feelings.

Table 5. Clinical Sample, Graded Model Item Parameter Estimates for six of the RFQY Scale B Items

Item	a	s.e.	b1	s.e.	b2	s.e.	b3	s.e.	b4	s.e.	b5	s.e.
11	0.49	0.11	-8.53	2.01	-7.66	1.78	-6.91	1.59	-3.51	0.80	0.21	0.20
18	1.11	0.13	-2.91	0.32	-1.83	0.21	-0.84	0.13	0.44	0.11	2.13	0.24
19	1.52	0.16	-2.67	0.25	-1.38	0.13	-0.69	0.10	0.31	0.09	1.76	0.17
34	2.04	0.22	-2.17	0.18	-1.38	0.12	-0.78	0.09	0.26	0.07	1.47	0.13
41	1.12	0.13	-3.57	0.41	-2.73	0.30	-2.00	0.22	-0.23	0.10	1.57	0.18
45	1.71	0.19	-2.65	0.24	-1.77	0.15	-1.29	0.12	-0.32	0.08	1.17	0.12

Table 6. Correlations between the RFQ-5 and experimental measures of mentalizing and self-report measures of emotion dysregulation and borderline traits (n = 100)

	MASC	CET	DERS						BPFSC	BPFSP
			Non-acceptance of emotions	Difficulties in goal-directed behavior	Impulse control difficulties	Lack of emotional awareness	Limited access to strategies	Lack of emotional clarity		
RFQY-5	0.150	.332**	0.020	-0.037	-.293**	-.618**	-0.039	-.320**	-.295**	-.283**
Scale B	0.127	.270**	0.132	0.007	-0.183	-.584**	0.033	-.279**	-.252*	-.251*

Notes. *p < .05; **p < .01. MASC = Movie Assessment of Social Cognition; CET = Children's Eyes Task; DERS = **Difficulties** in Emotion Regulation Scale; BPFSC = Borderline Personality Disorder Features Scale for Children.

Table 7. Group mean differences on versions of the RFQY between healthy controls and adolescent inpatients

	HC **Mean (SD)**	**Psychiatric** **Mean (SD)**	**df**	**t**	**p**	**d[95% CI]**
BPFSC	55.39 (14.38)	68.66 (15.35)	263	7.08	<.001	0.98[0.79, 1.16]
RFQ-5	4.43 (.82)	4.17 (.96)	280	-2.33	.020	0.25[.08, .42]
Scale B	4.29 (.47)	4.24 (.49)	280	-.93	.354	0.13[-.03, .31]

HC = Healthy controls