SAPIENZA
UNIVERSITÀ EDITRICE

Mentalizing Impairments, Pathological Personality and Aggression in Violent Offenders

Patrizia Velotti[a], Guyonne Rogier[a], Enrico Ciavolino[b], Paola Pasca[b], Susanne Beyer[a], Peter Fonagy[c]

[a]Department of Dynamic and Clinical Psychology, and Health Studies, Sapienza University of Rome, Rome, Italy
[b]Department of History, Society and Human Studies, Salento University, Lecce, Italy
[c]Psychoanalysis Unit, Research Department of Clinical, Educational and Health Psychology, University College London, London, United Kingdom

Psychology Hub (2021)
XXXVIII, 1, 51-60

Article info

Submitted: 02 February 2021
Accepted: 11 March 2021
DOI: 10.13133/2724-2943/17437

Abstract

Impairments in mentalizing abilities are thought to account for the high aggressive tendencies observed among individuals with pathological personality. However, the question of whether mentalizing impairments may mediate the pathways by which pathological personality leads to aggression has not yet been answered. This study first investigated the psychometric proprieties of the Italian version of the Reflective Functioning Questionnaire (RFQ). Then, we tested the mediating role of mentalizing in the relationship between the three pathological personality domains and aggressiveness. The study was conducted on a sample of 327 participants including a group of violent offenders (n=118) and a group of community participants (n=209). All subjects fulfilled the RFQ, the *Personality Inventory for DSM-5* (PID-5) and the *Aggression Questionnaire* (AQ). Partial Least Squares–Path Modelling with higher-order construct definition was used. Mentalizing capacities were shown to significantly mediate the pathways leading some pathological personality traits to aggression. Data supported the factorial structure of the RFQ found in the original validation study. Results also support the existence of a second-order variable, mentalizing, resulting from the convergence of hypomentalizing and hypermentalizing.

Keywords: mentalizing; RFQ; pathological personality; aggression, offenders.

*Corresponding author.
Patrizia Velotti
Department of Dynamic and Clinical
Psychology, and Health Studies,
Sapienza University of Rome,
Via degli Apuli, 1, 00185, Rome, Italy
Phone: + 39 06 44427556
E-mail: patrizia.velotti@uniroma1.it
(P. Velotti)

Introduction

As described within the psychoanalytic framework (Fonagy et al., 1993), one crucial aspect of human functioning consists of efforts to understand oneself and others. This arises from the consideration of intentions, thoughts, feelings, and needs in an interpersonal context in order to understand the self and also to anticipate others' behavior (Fonagy & Luyten, 2009). This reflective functioning ability—which has been termed *mentalizing* (Fonagy et al., 2016)—helps people to give sense to their emotional experiences. Recently, Luyten et al. (2020) stressed the multidimensional nature of the mentalizing construct. Indeed, this umbrella term is thought to include both automatic and controlled components as well as cognitive (e.g. perspective taking, belief-desire reasoning) and affective aspects (i.e. embodied mentalizing). In addition, the construct refers to a capacity to infer the nature of both own and others' mental states (Luyten et al., 2020). When individuals experience a strong disconnection between intentions, beliefs, thoughts, needs, and behaviors, they may fail to tolerate affects. In this situation, they endure a sense of the unknown and obscurity regarding their own and others' internal states, and can become overwhelmed by the intensity of affective experience. This is likely to result in increased impulsive behavior and a greater frequency of interpersonal conflicts.

Assessing mentalizing abilities is a complex issue. Only recently (Fonagy et al., 2016) an instrument assessing mentalizing capacity as a whole, the Reflective Functioning Questionnaire (RFQ), has been developed. Indeed, earlier research showed that mentalizing is a multifaceted construct, being the point of convergence of diverse abilities such as empathy, alexithymia, mindfulness, and theory of mind (Fonagy et al., 2011). Because several measures have been developed to assess definite constructs related to mentalizing, the authors have been mostly involved in the identification and assessment of specific facets of mentalizing.

There is evidence to support the hypothesis that individuals with an impairment in one of these abilities are likely to experience many adverse outcomes. For instance, a link between specific deficits in mentalizing abilities and a vulnerability for psychopathology (e.g., Bateman & Fonagy, 2004) has been shown. In addition, failures of mentalizing have been observed in individuals with borderline personality disorder (BPD), depression, and eating disorders (Fonagy & Luyten, 2016; Luyten & Fonagy, 2014). Difficulties related to mentalizing are thought to be a transdiagnostic feature across personality disorders that, however, can be distinguished as a function of the typical type of unbalance between the different polarities of mentalizing (Luyten et al., 2020). For instance, BDP would be associated with a deficit in controlled mentalizing resulting in an overreliance on automatic processes that, in turn, would lead to an impairment in cognitive mentalizing. A resulting proneness to excessive affective and other-focused mentalizing would explain most of the psychopathological functioning of individuals suffering from BPD (Luyten et al., 2020). Complementarily, recent contributions (Bateman et al., 2019; Simonsen & Euler, 2019) argue for the specific nature of mentalizing impairments among individuals suffering from narcissistic and avoidant personality disorder (switching rapidly to an automatic and affective mentalizing mode) or from antisocial personality disorder (suffering from serious impairments in affective mentalizing).

Moreover, these impairments are thought to account for the high aggressive tendencies observed among individuals with personality pathologies such as antisocial personality disorder (ASPD) or BPD (Bateman et al., 2016; Gillespie et al., 2018). Indeed, poor mentalizing-related capacities have been related to high levels of aggression (Euler et al., 2017). Unawareness of one's own affect (i.e., alexithymia) has been identified as a specific risk factor for aggression (Garofalo et al., 2018). Furthermore, poor mindfulness has recently been shown to account for aggressive tendencies in offenders (Velotti et al., 2016a; Velotti et al., 2018). Moreover, mindfulness, alexithymia, and empathy significantly interacted with aggression in predicting ASPD scores (Velotti et al., 2018).

Although a wide range of personality disorders has been observed among populations of offenders, some of them seem especially prevalent. For instance, BPD, Narcissistic Personality Disorder (NPD), and psychopathy are common diagnoses in the forensic field (Bernstein et al., 2007; Fazel & Danesh, 2002). However, the contrasting nature of results of empirical research into the links between personality disorders and aggression has been pointed out and explained in the light of the categorical approach adopted in these studies. Indeed, in the field of personality disorders, comorbidity is more common than not, and some authors stressed the utility of adopting a dimensional perspective to reach an optimal understanding of personality disorder (Widiger & Trull, 2007; Widiger & Simonsen, 2005). Following this line, the most recent version of the DSM (American Psychiatric Association, 2013) proposed a dimensional approach to personality disorder, identifying five main domains of pathological personality: Negative Affect, Detachment, Antagonism, Disinhibition, and Psychoticism. Although all these domains may to some extent account for greater proneness to aggression in both clinical and normal populations (Dunne et al., 2018), three of them appear to assume a potential stronger explanatory role of the mechanisms that lead to aggressive behaviors among offenders. First, Negative Affect converges in the description of BPD and NPD (Wright et al., 2013), including facets that have previously been reported to be significantly associated with aggression proneness, such as emotional lability (Dvorak et al., 2013; Velotti et al., 2017), separation insecurity (Critchfield et al., 2008), and hostility (e.g., Jones et al., 2011). Second, the Disinhibition domain is a reasonable candidate in explaining aggressive behavior, as it encompasses personality facets that have been found to account for aggression. For instance, a number of studies have provided evidence for the role of Impulsivity in aggression (Bettencourt et al., 2006; Garofalo et al., 2018), as well as Risk-Taking (Miller et al., 2012). Finally, the Deceitfulness, Manipulativeness, Callousness, and Grandiosity facets of the Antagonism domain are central descriptors of two personality disorders widely observed among offenders—NPD and psychopathy (Bettencourt et al., 2006; Enebrink et al., 2005; Jones et al., 2011).

Despite these insightful contributions focusing on partial facets of the construct of mentalizing, a comprehensive picture of the role played by mentalizing impairments in aggression is still lacking. Notwithstanding the recent development of the RFQ, to our knowledge, no study has assessed mentalizing as a unitary construct in a population of community participants and offenders. Moreover, the question of whether mentalizing impairments may mediate the pathways by which pathological personality leads to aggression has not yet been answered.

On the basis of these considerations, the present study aimed to evaluate the psychometric proprieties of the Italian version of the RFQ in both community participants and offenders and to test, among a population of offenders, its mediating role in the relationship between specific domains of pathological personality (Negative Affect, Disinhibition, and Antagonism) and aggression.

Method

Participants

The study was conducted on a sample of 327 participants divided into two groups. The offender group (*M*age = 38.72; *SD* = 11.75) was composed of 118 Italian men convicted for violent crimes and recruited from prisons in the Latium and Liguria regions. These participants have been recruited throughout a convenience sampling method. Specifically, participation to the study was proposed to all individuals convicted for violent crimes by operators of educational services operating in the single jails. In case of interest, an appointment with a member of the research team was organized. The control group comprised 209 males recruited from the general population by a purposive sampling method. Specifically, students in psychology courses were asked to recruit community participants. The mean age of this group was 41.85 years (*SD* = 13.64). As the RFQ missing values represented the 20% or less of the total observations, data were imputed to the mean.

Procedure

Before their involvement in the study, all the participants were informed about the study's aim and scope. Information about anonymity and privacy was provided, and participants were asked to provide written informed consent. Then, a battery of self-report questionnaires was administered under the supervision of a psychologist. Noteworthy, community participants were only asked to fulfill the demographic information questionnaire and the RFQ (detailed below) whereas an extended version of the battery was administered to the offender group. The study received formal approval from the Research Ethics Board of the University of Rome, Sapienza and the Italian Ministry of Justice.

Measures

A *Demographic information questionnaire*, expressly created for the purposes of this study, was used to collect information as such as age and gender.

The *Reflective Functioning Questionnaire* (RFQ; Fonagy et al., 2016) is an 8-item self-report questionnaire scored on a 7-point Likert scale ranging from 1 (*Strongly disagree*) to 7 (*Strongly agree*). The instrument provides two scores referring to two subscales, labeled *Certainty about Mental States* (RFQ_C), measuring hypermentalizing, and *Uncertainty about Mental States* (RFQ_U), measuring hypomentalizing. Instructions for scoring were retrieved from the developer's website and consist in recoding and summing six items of the instrument to obtain the total RFQ_C score and recoding and summing six items (partially overlapping with those used in the scoring of the RFQ_C) to obtain the RFQ_U score. In our study, internal consistencies of each subscale were in line with the original validation study (Fonagy et al., 2016), with Cronbach's alphas of .667 for the RFQ_U subscale and .782 for the RFQ_C. The official Italian version of the instrument, performed following a back-translation procedure, was retrieved from the developers' website.

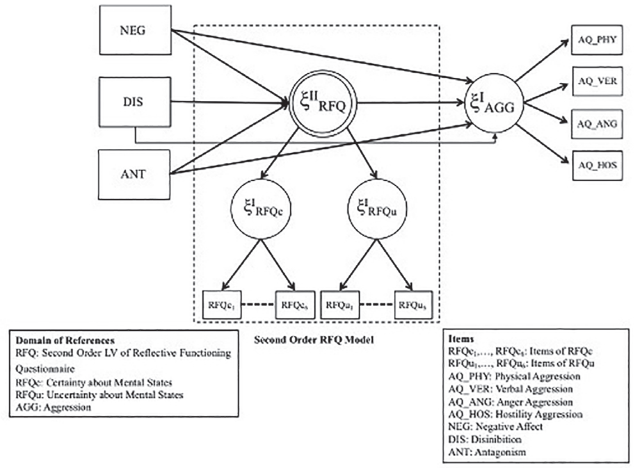
The *Aggression Questionnaire* (AQ; Buss & Perry, 1992; Fossati et al., 2003) is a 29-item multidimensional self-report scale used to assess trait aggression. Participants belonging to the offender's group are asked to answer on a 5-point Likert scale ranging from 1 (*Extremely uncharacteristic of me*) to 5 (*Extremely characteristic of me*). The four dimensions of the AQ converge to provide a total score measuring levels of aggression. In the present study, the AQ was confirmed to have excellent psychometric proprieties, with an internal consistency of the AQ total score of .90.

The *Personality Inventory for DSM-5* (PID-5; Krueger et al., 2013; Fossati et al., 2013) is a self-report questionnaire encompassing 220 items scored on a 4-point Likert scale ranging from 0 (*Very false or Often false*) to 3 (*Very true or Often true*). Only participants belonging to the offender's group fulfill this self-report questionnaire. It assesses pathological personality and provides a score for each of the 25 subscales, which measure maladaptive personality facets. Moreover, some of these facets converge in five mains domains of pathological personality. The *Negative Affect* domain describes an individual likely to experience negative emotions, anxiety, and separation insecurity. The *Antagonism* domain refers to aggressive tendencies of dominance, grandiosity, and deceitfulness. The *Psychoticism* domain evaluates disconnection from reality, eccentricity, unusual beliefs and experiences, and cognitive dysregulation. The *Detachment* domain indicates a proneness to social isolation, anhedonia, and avoidance. Finally, the *Disinhibition* domain includes impulsivity and sensation-seeking proneness. In our study, the psychometric properties of the instrument were confirmed, with all Cronbach's alphas being ≥ .88.

Statistical Analyses

When it comes to theoretical development and improvement, psychology traditionally relies on parametric factorial analytic approaches to modeling relationships between unobservable constructs, even in spite of the absence of preliminary, mandatory requirements (e.g. a well-structured research design, distributional assumptions of the data; Fabrigar, Wegener, MacCallum, & Strahan, 1999). Recently, Structural Equation Modeling (SEM) based on Partial Least Squares (PLS) became a valid alternative or complement and is being more widely used (Ciavolino & Nitti, 2013a; 2013b), as it represents the non-parametric alternative to Covariance-Based SEM (CB-SEM) (Ciavolino & Nitti, 2013a; 2013b; Lohmöller, 1989; Nitti & Ciavolino, 2014). As it provides constructs approximations, it is conceptually closer to the psychological context (Hair, Hult, Ringle, & Sarstedt, 2016). However, as Rönkkö, McIntosh, and Antonakis, (2015) pointed out, the classic PLS approach presents some flaws (e.g. increased risk of false positives, lack of formal assessment and guidelines) whereas the CB-SEM, with its more strict criteria, yields more accurate results. In order to improve PLS-SEM, Dijkstra and Henseler (2015) developed a PLS variant providing corrected estimates and, most of all, reducing the false positive risk within the structural model. PLSc is currently being willingly accepted from the research community, as it is in close correspondence with the CB-SEM (Aguirre-Urreta, & Rönkkö, 2018; Rönkkö, McIntosh, & Aguirre-Urreta, 2016). In light of the sample characteristics, a research design which would limit researchers in adopting CB-SEM, and also keeping into account the latest statistical improvements, we opted for the PLS-SEM approach, using the PLS consistent variant to define and analyse the theoretical model.

Fig. 1. Theoretical model tested


NEG
DIS
ANT
ξ^II RFQ
ξ^I AGG
ξ^I RFQc
ξ^I RFQu
AQ_PHY
AQ_VER
AQ_ANG
AQ_HOS
RFQc1
RFQc6
RFQu1
RFQu6
Domain of References
RFQ: Second Order LV of Reflective Functioning Questionnaire
RFQc: Certainty about Mental States
RFQu: Uncertainty about Mental States
AGG: Aggression
Second Order RFQ Model
Items
RFQc1,…, RFQc6: Items of RFQc
RFQu1,…, RFQu6: Items of RFQu
AQ_PHY: Physical Aggression
AQ_VER: Verbal Aggression
AQ_ANG: Anger Aggression
AQ_HOS: Hostility Aggression
NEG: Negative Affect
DIS: Disinibition
ANT: Antagonism

Consistent with the theoretical path model (illustrated in Figure 1), the following manifest and latent variables were defined:

1) Manifest Variables (MVs): Negative Affect (NEG), Disinhibition (DIS), and Antagonism (ANT), according to the theoretical model, were the MVs that can have a direct and indirect effect on the LV Aggression. Moreover, AQP (Aggression Questionnaire Physical), AQV (Aggression Questionnaire Verbal), AQA (Aggression Questionnaire Anger), and AQH (Aggression Questionnaire Hostility) are MVs measuring the LV Aggression.
2) First-order LV: The first-order LV is Aggression, concerning the mediation variable.
3) Second-order LV: RFQ refers to the Reflective Functioning Questionnaire.

The first step of the statistical analyses involved the assessment of the psychometrics of the RFQ scale.

The hierarchical levels of the RFQ scale were analysed through the *Hierarchical Component Model* or *Repeated Indicators Approach* (RIA; Lohmöller, 1989), where the manifest indicators of each first-order LV are simply repeated at each level of the hierarchy to represent the higher-order constructs.

Especially when it comes to measures that assess opposite sub-dimensions, it may be the case that the corresponding indicators would show opposite loadings (meaning that they also display a negative correlation), which is not desirable, considering a scale assessing just one construct. In view of this issue, Sanchez (2013) suggests changing the sign of half of the indicators (in this case, either the *Certainty* or the *Uncertainty* subscales of the RFQ). In our study, we chose to do this for the *Uncertainty* subscale: once the sign is inverted, it can be said that the subscale measures the *Lack of Uncertainty* about mental states. Cronbach's alpha, Dillon-Goldstein's rho (DG-rho) and composite reliability indices assessed indicators (i.e., MVs) reliability and their coherence with the respective construct (i.e., the same LV). At the same time, measurement invariance among *offenders* and *non offenders* has been tested: it is of fundamental importance to establish that items are perceived in the same manner by different groups (Hair, et al., 2016).

After a panoramic view of the psychometrics of the RFQ scale, *Reflective Functioning* was modeled via the *two-steps* approach (Tenenhaus, 2008) and embedded into a more complex relationship model, in which it acts as a mediator. Scores of the previously analyzed dimensions *Certainty* and *Lack of certainty* defined the higher-order construct *Reflective Functioning*. Reliability and validity of the model have been evaluated: in particular, convergent validity was assessed via inspection of the loadings and the heterotrait-monotrait (HTMT) ratio of correlations (Hair Jr, Hult, Ringle, & Sarstedt, 2016; Rönkkö, & Cho, 2020). While loadings inform us about the correlation between LVs and their indicators, the HTMT matrix expresses the correlation between LVs, if they were perfectly measured. As the typical assumptions of the PLS approach does not allow to use parametric significance tests for the estimated coefficients, researchers relied on a non-parametric bootstrap procedure (with 5000 samples, as suggested by the current literature, Hair et al., 2016; Tenenhaus, et al., 2005), which provided a better approximation of the estimated coefficients, allowing insights on bias, standard errors, as well as on significance. Finally, the relationship between the dimensions of the RFQ and other clinical variables was explored by the calculation of Pearson's *r* correlations.

Results

Scale Evaluation

RFQ evaluation

For each of the RFQ subscales, Table 1 reports the loadings of the global sample (N = 327 of which n = 118 *Offenders* and n = 209 *non Offenders*) computed via PLS consistent algorithm. Although some items (taken singularly) showed low loadings, they were retained for several, theoretical reasons: first, the item RFQ_u2 intrinsically characterized the respective sub dimension: *I always* *know what I feel* reflects the lack of uncertainty about one's own feelings; the same can be said for the RFQc1, *People's thoughts are a mistery to me*. Second, the current literature (Badoud et al., 2015) considers the aggregated items (by either mean or median) forming each subdimension of the RFQ scale.

Tab. 1. Loadings and cross-loadings of the RFQ subscales

	RFQc	NRFQu
NRFQu2	0.099	**0.200**
NRFQu3	0.394	**0.493**
NRFQu5	0.445	**0.599**
NRFQu6	0.502	**0.591**
NRFQu7	0.442	**0.546**
NRFQu8	0.451	**0.590**
RFQc1	**0.322**	0.182
RFQc3	**0.594**	0.416
RFQc4	**0.722**	0.587
RFQc6	**0.641**	0.513
RFQc7	**0.658**	0.51
RFQc8	**0.745**	0.637

In general, the sub-dimensions of the RFQ show a sufficient/good reliability, as table 2 shows.

Tab. 2. Reliability of the RFQ scale

	Cronbach's α	Dillon-Goldstein's ρ	Composite reliability	AVE
RFQ	0.816	0.844	0.824	0.295
RFQc	0.782	0.812	0.789	0.396
RFQu	0.667	0.706	0.675	0.272

All the indexes range between 0.7 and 0.9 even though the AVE, that is, the ratio between the squared standardized loadings and the number of manifest variables, suggests the constructs explain less than the 50% of the variance of the indicators. In line with the recent literature on PLS-SEM, measurement invariance can be tested through a 3-step permutation-based procedure, the *Measurement invariance of composite models* (MICOM, Henseler, Ringle, & Sarstedt, 2016). As researchers engage in testing measurement invariance, it is important to ensure that the number of observations in each group meets the rules of thumb for minimum sample size requirement (Hair Jr, Sarstedt, Ringle, & Gudergan, 2017). Considering a statistical power of the 80% and a significance level α = 0.05, Hair Hult, Ringle, and Sarstedt's (2016) recommendation, drawn from and adapted by Cohen's (1992), would suggest a minimum of 90 participants per group, in order to observe an R^2 of at least 0.10 in any endogenous variable (in particular, $R^2 > 0.8$ for RFQc and $R^2 > 0.7$ for RFQu in both *Offenders* and *non Offenders*). Requirement being met in the present study. The research design granted *configural* invariance (the first step), as indicators were the same across each group's measurement model, and groups themselves were treated equally (e.g. reverse items, data coding). Table 3 reports the results of the MICOM procedure (1000 permutations),

Step 2 assesses whether the composite scores significantly differ across the groups: the null hypothesis tells that *c*, that is, the correlations between the composite scores, should be 1; compositional invariance would be confirmed if *H* were not rejected. The final step (step 3) computes the permutation scores of the aggregated data and tests whether means and variances differ significantly. Full measurement invariance (and therefore, the possibility to either run multigroup analyses or to use the pooled dataset for research conclusions) is met if those differences are not statistically significant from one another. As it can be noted from the *p*-values of step 2, as well as from the confidence intervals reported in the table, all including 0, the RFQ shows full measurement invariance.

Theoretical model evaluation

Subsequently, the *Offenders* group allowed researchers to consider *Reflective Functioning* (given by the scores of both RFQ sub-dimensions) construct as a mediator within a more complex relationship system. As shown in Table 4, for the LVs *RFQ* and *Aggression*, both the average inter-variable correlation α, the DG-rho and the composite reliability index are higher than 0.7, meaning that the LVs are well represented by their respective indicators.

Tab. 4. Dillon-Goldstein's rho and Cronbach's α indices (n=118)

	Mode	MVs	Cronbach's α	Dillon-Goldstein's ρ	Composite reliability	AVE
RFQ	A	2	0.759	0.762	0.760	0.613
Aggression	A	4	0.863	0.882	0.862	0.616

Note. MVs: Manifest Variables; RFQ: Reflective Functioning Questionnaire.

Tab. 3. Results of the MICOM procedure

	Step 2. Compositional Invariance			Step 3 Equal Means and Variances					
LV	c	5.00% quantile of cu	Permutation p-value	Mean difference	CI (lower)	CI (upper)	Variance difference	2.50%	97.50%
RFQ	**0.997**	0.992	**0.451**	**0.055**	-0.211	0.218	**0.005**	-0.298	0.347
RFQc	**0.997**	0.993	**0.299**	**-0.096**	-0.209	0.234	**0.134**	-0.231	0.253
RFQu	**0.992**	0.978	**0.352**	**0.234**	-0.228	0.227	**-0.135**	-0.586	0.667

Convergent Validity

The confidence intervals of the indicator loadings in Table 5 all reflect significant correlations with their one and only corresponding LV. The HTMT matrix (see Table 6) suggests, instead, well-distinct constructs and therefore a good discriminant validity: in fact, all the ratios are lower than the most conservative threshold of 0.85 (Henseler, & Sarstedt, 2013). Both *Certainty* and *Uncertainty* about one's own mental states summarize *Reflective Function*, while *Aggression* is well represented by the indicators *AQ_Physical*, *AQ_Verbal*, *AQ_Anger*, and *AQ_Hostility*. We can conclude that all the indicators properly measure their respective constructs, thus confirming the discriminant validity of the scales.

Tab. 5. Loadings and confidence intervals (n=118)

	Original	Mean Bootstrap	SE	CI 2.5th percentile	CI 97.5th percentile
RFQ–RFQ_C	0.813	0.814	0.029	0.749	0.864
RFQ–RFQ_U	0.752	0.751	0.038	0.663	0.815
Aggress–AQ_PHY	0.808	0.807	0.075	0.629	0.925
Aggress–AQ_VER	0.564	0.561	0.107	0.322	0.741
Aggress–AQ_ANG	0.821	0.819	0.056	0.691	0.911
Aggress–AQ_HOS	0.904	0.895	0.062	0.787	1.030

Note. All the coefficients are statistically significant given the bootstrap results (1000 replications).

Tab. 6. Heterotrait Monotrait Ratio (n=118)

	Aggress	Antagonism	Dishinibition	Negative_Affect	RFQ
Aggress					
Antagonism	0.509				
Dishinibition	0.570	0.532			
Negative_ Affect	0.497	0.662	0.649		
RFQ	0.711	0.430	0.520	0.560	

Note. Values represent a deattenuated correlation. In other words, an estimate of the correlations between the constructs if they were perfectly measured.

Structural Model

A first evaluation of the structural model in the PSL-SEM context can be provided by the coefficient of determination, R^2, which represents the amount of variance in the endogenous LV explained by its independent LV. Table 4 reports the results of this assessment. For *RFQ*, the R^2 score is low *(0.343)* but acceptable and statistically significant *($p < 0.01$)*. On the other side, the *Aggression* scale for the LVs shows a moderate amount of explained variance *($R^2 = 0.554, p < 0.001$).*

Differently from the LVs *Negative Affect*, which showed an almost null effect size on *Aggression* *($f^2 = 0.015$).*, *Antagonism* and RFQ showed a small and a large impact on it, respectively *($f^2 = 0.092$* and *$f^2 = 0.571$)*. A summary of the bias-corrected estimated path coefficients, along with the *t* statistics and significance of both direct and indirect effects is reported in Table 7 (bootstrap with 5000 replications).

As is can be noted, *Antagonism* directly and significantly impacts on *Aggression*. Other LVs such as *Negative Affect* and *Dishinibition* directly, negatively and significantly impact on *Reflective Functioning*. Increased *Dishinibition*, as well as increased *Negative Affect* seem to be linked to lower *Reflective Functioning*.

In terms of direct effect, a strong negative and significant impact of the *Reflective Functioning* on *Aggression* can be noted as well. The more one is conscious about his own mental state, the less tends to aggressiveness. Considering *Reflective Functioning* as a mediator, two out of three mediation emerge as statistically significant. In other words, *Negative Affect* may lay the ground for *Aggression* only when there is a lack of *Reflective Functioning*.

Correlational analyses

In the group of offenders, we performed correlational analyses to explore the relationships between the RFQ subscales and both aggression and pathological personality. Results, displayed in Table 8, showed that the RFQ_U subscale correlates

Tab. 7. Bias-corrected estimates, *t* statistics and confidence interval of direct and indirect effects (n=118)

Direct Effects	*β* Original	*β* Bootstrap	*SD*	Distorsion	t	CI 2.5%	CI 97.50%	f ²	CI 2.5%	CI 97.50%
ANT -> AGGRESS	0.310	0.310**	0.103	0.000	3.021	**0.091**	**0.491**	0.092	0.003	0.338
ANT -> RFQ	-0.068	-0.073	0.135	-0.005	0.505	-0.338	0.200	0.004	0.000	0.102
DIS -> AGGRESS	0.158	0.160	0.082	0.002	1.934	0.026	0.346			
DIS -> RFQ	-0.261	-0.256*	0.111	0.005	2.353	**-0.478**	**-0.042**	0.060	0.001	0.238
NEG -> AGGRESS	0.195	0.191	0.102	-0.004	1.910	-0.005	0.397	0.000	0.000	0.085
NEG -> RFQ	-0.346	-0.340**	0.128	0.006	2.698	**-0.596**	**-0.095**	0.082	0.004	0.283
RFQ -> AGGRESS	-0.605	-0.613***	0.124	-0.008	4.886	**-0.832**	**-0.352**	0.571	0.165	1.931

Indirect Effects	*β* Original	*β* Bootstrap	*SD*	Distorsion	t	CI 2.5%	CI 97.50%
ANT -> RFQ -> AGGRESS	0.041	0.046	0.085	0.005	0.482	-0.115	0.228
DIS -> RFQ -> AGGRESS	0.158	0.160	0.082	0.002	1.934	**0.026**	**0.346**
NEG -> RFQ -> AGGRESS	0.210	0.209*	0.092	-0.001	2.272	**0.062**	**0.434**

Note. The path coefficients and the confidence intervals are estimated by the bootstrap procedure with 200 replications. Confidence intervals for significant coefficients (in bold) are those that do not contain zero; NEG: Negative Affect Domain of the PID-5; RFQ: Reflective Functioning Questionnaire; AGGRESS: Aggression; DIS: Disinhibition Domain of the PID-5; ANT: Antagonism domain of the PID-5.

positively and significantly with aggression and pathological personality measures, whereas the RFQ_C subscale shows the reverse pattern of results.

Discussion

This study aimed to assess reflective functioning in offenders in order to evaluate its predictive role on aggression as well as its mediating role in the pathway that leads from pathological personality to aggression. To achieve this goal, we first assessed the psychometric properties of the RFQ. Data supported the factorial structure found in the original study (Fonagy et al., 2016). Results also support the existence of a second-order variable, mentalizing, resulting from the convergence of hypomentalizing and hypermentalizing. In addition, we found good internal consistency indices in our sample of offenders, extending the existing literature on the psychometric properties of the RFQ.

The PID-5 domains positively correlated with aggression levels. This finding is congruent with theoretical assumptions and empirical evidence linking personality disorders to high aggressive tendencies (Bernstein et al., 2007; Fazel & Danesh, 2002).

In addition, among offenders, the RFQ subscales correlated with all pathological personality domains investigated. The highest association between the RFQ_U subscale and the PID-5 domains was found for the *Negative Affect* domain. This domain is primarily related to BPD (Calvo et al., 2016), further supporting the hypothesis of a tight relationship between impairments in mentalizing and borderline personality features.

However, the two RFQ subscales showed an inverse pattern of associations, with RFQ_C being negatively associated, and RFQ_U positively associated, with pathological personality. In this context, it appears that the role of RFQ_C in psychopathology is neither adaptive nor maladaptive *per se*. This converges with the study of Cucchi, Hampton and Moulton-Perkins (2018) that, for instance, suggests that hypermentalizing levels do not characterize clinical sample with aggression towards the self. It could be speculated that extreme levels of *Certainty about mental states* are not typical of Axis II disorders, extending the previous literature regarding BPD features (Badoud et al., 2015; Fonagy et al., 2016).

The subscales of the RFQ showed the same pattern of results in relation to measures of aggression. These data are in line with a wide range of indirect evidence showing relationships between focused mentalizing impairments (e.g., in theory of mind, empathy, or alexithymia) and aggressiveness among offenders (Velotti et al., 2016b, 2018). This supports the hypothesis that the capacity to understand one's own and others' minds plays a key role in aggressive behavior (Bateman et al., 2013; Tolan et al., 2013).

Contrary to our hypothesis, we observed that average means of mentalizing levels did not differ between offenders and community participants. On one hand, this suggests that offenders convicted for violent crimes are not characterized by poor reflective functioning. On the other hand, we may speculate that other moderating variables, not examined in our study, may explain this result. For instance, we did not differentiate between the type of violent crime committed by our sample of offenders, but we could formulate the hypothesis that motivations underlying the illegal act may discriminate between offenders with poor and normal mentalizing capacities. From this perspective, the replication of our study keeping in mind the distinction established by Meloy (Hoffer et al., 2018) between predatory and affective violence would probably bring a precious additional insight regarding this issue.

Regarding the mediational analyses, interesting results emerged. First, we found that the relationship between *Negative Affect* and aggression was entirely mediated by levels of mentalizing. This is in line with theoretical literature asserting that individuals with BPD would show aggression because of a central impairment in their mentalizing function that compromises their emotion-regulation capacities. Regarding the *Disinhibition* domain, a more nuanced picture emerged, indicating that the pathway leading this pathological domain to aggression was only partially mediated by levels of mentalizing. This result suggests that although some facets of this pathological domain—such as impulsivity and risk-taking—are reasonably related to mentalizing impairments, others—such as a lack of rigid perfectionism or irresponsibility—may account for aggression through alternative pathways that involve, for instance, social and cultural aspects. Finally,

Tab. 8. Correlations between mentalizing functions, pathological personality, and aggression (n=118)

		RFQ				PID-5				AQ
		Gen	Cer	Unc	Psy	Ant	Dis	Det	Neg	Total
RFQ	General	–								
	Cer	.91**	–							
	Unc	.88**	–.52**	–						
PID-5	Psy	-.53**	–.58**	.40**	–					
	Ant	-.38**	–.46**	.19*	.59**	–				
	Dis	-.39**	–.46**	.35**	.70**	.53**	–			
	Det	-.48**	–.38**	.33**	.65**	.53**	.85**	–		
	Neg	-.49**	–.47**	.39**	.71**	.66**	.64**	.64**	–	
AQ	Total	-.58**	–.58**	.42**	.43**	.49**	.53**	.41**	.47**	–

Note: RFQ: Reflective Functioning Questionnaire; PID-5: Personality Inventory for DSM-5; AQ: Aggression Questionnaire; Cer: Certainty; Unc: Uncertainty; Psy: Psychoticism; Ant: Antagonism; Dis: Disinhibition; Det: Detachment; Neg: Negative Affect; **p* < .05; ***p* < .001.

and in contrast to our initial hypothesis, we found that the relationship between *Antagonism* and aggression was not mediated by levels of mentalizing. It should be noted that, this domain is especially linked to narcissistic and psychopathic traits that may be characterized by instrumental aggressiveness rather than emotionally disrupting behaviors (Meloy, 2006).

Limitations and future directions

Our study has important strengths, such as the innovative use of the RFQ and PID-5 in a sample of offenders. However, it is not without limitations. First, the use of self-report questionnaires may be inconsistent with the construct we are aiming to scrutinize, and some may question the validity of our both our self-report variables in this context. Future research should perhaps also make use of more sensitive interview-based approaches to assess the severity of variables such as personality and reflective function. Second, some potential moderators of the relationship between reflective functioning and aggression have not been examined. For instance, the role of cognitive functioning or specific facets of mentalizing (e.g., alexithymia, empathy, perspective-taking) as moderators needs to be further explored. Finally, future studies should analyze the role of mentalizing in relation to self-directed aggression in order to form a complete picture of the complex interplay between mentalizing abilities and aggression.

Author Contributions

The authors contributed equally to this manuscript.

Compliance with Ethical Standards

Conflict of interest

The authors declare that they have no competing interests.

Funding

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

Ethical approval

This study was approved by the Ethics Board of the Department of Dynamic and Clinical Psychology, and Helath Studies, University of Rome, Sapienza.

Availability of data and material

The data that support the findings of this study are available from the corresponding author [PV], upon reasonable request.

References

Aguirre-Urreta, M. I., & Rönkkö, M. (2018). Statistical inference with PLSc using bootstrap confidence intervals. *MIS quarterly*, *42*(3), 1001-1020.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC.

Badoud, D., Luyten, P., Fonseca-Pedrero, E., Eliez, S., Fonagy, P., & Debbané, 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): e0145892. https://doi.org/10.1371/journal.pone.0145892

Bateman, A., Bolton, R., & Fonagy, P. (2013). Antisocial personality disorder: A mentalizing framework. *FOCUS: The Journal of Lifelong Learning in Psychiatry*, *11*(2), 178–186. doi: 10.1176/appi.focus.11.2.178

Bateman, A., & Fonagy, P. (2004). Mentalization-based treatment of BPD. *Journal of Personality Disorders*, *18*(1), 36–51. doi: 10.1521/pedi.18.1.36.32772

Bateman, A., O'Connell, J., Lorenzini, N., Gardner, T., & Fonagy, P. (2016). A randomised controlled trial of mentalization-based treatment versus structured clinical management for patients with comorbid borderline personality disorder and antisocial personality disorder. *BMC Psychiatry*, *16*(1), 304. doi: 10.1186/s12888-016-1000-9

Bateman, A., Motz, A., & Yakeley, J. (2019). Antisocial personality disorder in community and prison settings. In Bateman, A., & Fonagy, P., (2019). *Handbook of Mentalizing in Mental Health Practice*. Am. Psychiatr. Publ. 2nd ed.

Bernstein, D. P., Arntz, A., & de Vos, M. (2007). Schema focused therapy in forensic settings: Theoretical model and recommendations for best clinical practice. *International Journal of Forensic Mental Health, 6*, 169–183. doi: 10.1080/14999013.2007.10471261

Bettencourt, B. A., Talley, A., Benjamin, A. J., & Valentine, J. (2006). Personality and aggressive behavior under provoking and neutral conditions: A meta-analytic review. *Psychological Bulletin*, 132, 751–777. doi: 10.1037/0033-2909.132.5.751

Buss, A.H., & Perry, M. (1992). The Aggression Questionnaire. *Journal of Personality and Social Psychology*, 63, 452–459.

Calvo, N., Valero, S., Sáez-Francàs, N., Gutiérrez, F.,Casas, M. & Ferrer, M. (2016). Borderline Personality Disorder and Personality Inventory for DSM-5 (PID-5): Dimensional personality assessment with DSM-5. *Comprehensive Psychiatry, 70,* 105–111. doi: 10.1016/j.comppsych.2016.07.002

Ciavolino, E., & Nitti, M. (2013). Using the Hybrid Two-Step estimation approach for the identification of second-order latent variable models. *Journal of Applied Statistics*. Volume 40, Issue 3, 508-526.

Ciavolino, E., & Nitti, M. (2013). Simulation study for PLS path modelling with high-order construct: A job satisfaction model evidence. In: Proto A., Squillante M., Kacprzyk J. (eds) *Advanced Dynamic Modeling of Economic and Social Systems. Studies in Computational Intelligence*. Volume 448, Springer, Berlin Heidelberg, 185-207.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Critchfield, K. L., Levy, K. N., Clarkin, J. F., & Kernberg, O. F. (2008). The relational context of aggression in borderline personality disorder: Using adult attachment style to predict forms of hostility. *Journal of Clinical Psychology, 64*, 67–82. doi: 10.1002/jclp.20434

Cucchi, A., Hampton, J.A., & Moulton-Perkins, A. (2018). Using the validated Reflective Functioning Questionnaire to investigate mentalizing in individuals presenting with eating disorders with and without self-harm. *PeerJ* 6:e5756. Doi:10.7717/peerj.5756

Dijkstra, T. K., & Henseler, J. (2015). Consistent partial least squares path modeling. *MIS quarterly*, *39*(2), 297-316.

Dunne, A. L., Gilbert, F., & Daffern, M. (2018). Investigating the relationship between DSM-5 personality disorder domains and facets and aggression in an offender population using the Personality Inventory for the DSM-5. *Journal of Personality Disorders*, *32*(5), 668–693. doi: 10.1521/pedi_2017_31_322

Dvorak, R. D., Pearson, M. R., & Kuvaas, N. J. (2013). The five-factor model of impulsivity like traits and emotional lability in aggressive behavior. *Aggressive Behavior, 39*, 222–228. doi: 10.1002/ab.21474

Enebrink, E., Andershed, H., & Langstrom, N. (2005). Callous-unemotional traits are associated with clinical severity in referred boys with conduct problems. *Nordic Journal of Psychiatry*, 59, 431–440. doi: 10.1080/08039480500360690

Euler, F., Steinlin, C., & Stadler, C. (2017). Distinct profiles of reactive and proactive aggression in adolescents: Associations with cognitive and affective empathy. *Child and Adolescent Psychiatry and Mental Health, 11*, 1. doi: 10.1186/s13034-016-0141-4

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272.

Fazel, S., & Danesh, J. (2002). Serious mental disorder in 23000 prisoners: A systematic review of 62 surveys. *Lancet*, 359(9306), 545–550. doi: 10.1016/S0140-6736(02)07740-1

Fonagy, P., Bateman, A., & Bateman, A. (2011). The widening scope of mentalizing: A discussion. *Psychology and Psychotherapy: Theory, Research and Practice, 84*, 98–110. doi: 10.1111/j.2044-8341.2010.02005.x

Fonagy, P., & Luyten, P. (2009). *Mentalization and borderline personality disorder*. Oxford University Press.

Fonagy, P., & Luyten, P. (2016). A multilevel perspective on the development of borderline personality disorder. In: D. Cicchetti (Ed), *Developmental psychopathology. Vol 3: Maladaptation and psychopathology. 3rd ed.* (pp. 726–792). New York, NY: John Wiley & 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., Moran, G. S., & Target, M. (1993). Aggression and the psychological self. *Praxis Der Kinderpsychologie Und Kinderpsychiatrie*, *47*(3), 125–143.

Fossati, A., Kreuger, R. F., Markon, K. E., Borroni, S., & Maffei, C. (2013). Reliability and validity of the Personality Inventory for DSM-5 (PID-5). *Assessment, 20*(6), 689–708. doi: 10.1177/1073191113504984

Fossati A., Maffei C., Acquarini E., & Di Ceglie A. (2003). Multigroup confirmatory component and factor analyses of the Italian version of the Aggression Questionnaire. *European Journal of Psychological Assessment, 19,* 54–65. doi: 10.1027//1015-5759.19.1.54

Garofalo, C., Velotti, P., & Zavattini (2018). Emotion regulation and aggression: The incremental contribution of alexithymia, impulsivity, and emotion dysregulation facets. *Psychology of Violence, 8*(4), 470–483. doi: 10.1037/vio0000141

Gillespie, S., Garofalo, C., Velotti, P. (2018). Emotion regulation, mindfulness, and alexithymia: Specific or general impairments in sexual, violent, and homicide offenders? *Journal of Criminal Justice, 58,* 56-66. doi: 10.1016/j.jcrimjus.2018.07.006

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

Hair Jr, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2017). *Advanced issues in partial least squares structural equation modeling*. Sage publications.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *International marketing review*. *33*(3), 405-431.

Henseler, J., & Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modeling. *Computational statistics*, *28*(2), 565-580.

Hoffer, T., Hargreaves-Cormany, H., Muirhead, Y., Meloy, J.R. (2018) Meloy's Bimodal Theory of Affective (Reactive) and Predatory (Instrumental) Violence. In: *Violence in Animal Cruelty Offenders*. SpringerBriefs in Psychology. Springer, Cham. https://doi.org/10.1007/978-3-319-91038-3_7

Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, anti-social behavior, and aggression: A meta-analytic review. *Journal of Criminal Justice*, 39, 329–337. doi: 10.1016/j.jcrimjus.2011.03.004

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2013). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological Medicine, 42*, 1879–1890. doi: 10.1017/S0033291711002674

Lohmöller, J.B. (1989). *Latent Variable Path Modeling with Partial Least Squares*, Physica-Verlag, Heidelberg.

Luyten, P., Campbell, C., Allison, E., & Fonagy, P. (2020). The Mentalizing Approach to Psychopathology: State of the Art and Future Directions. *Annual review of clinical psychology, 16*, 297–325. https://doi.org/10.1146/annurev-clinpsy-071919-015355

Luyten, P., & Fonagy, P. (2014). Psychodynamic treatment for borderline personality disorder and mood disorders: a mentalizing perspective. In: L. Choi-Kain, & J. Gunderson (Eds). *Borderline personality disorder and mood disorders: Controversies and consensus* (pp. 223–251). Springer.

Meloy, J.R., (2006). Empirical basis and forensic application of affective and predatory violence; *Australian and New Zealend journal of psychiatry*, 40:539-547

Miller, J. D., Zeichner, A., & Wilson, L. F. (2012). Personality correlates of aggression: Evidence from measures of the Five-Factor Model, UPPS model of impulsivity, and BIS/BAS. *Journal of Interpersonal Violence,* 14, 2903–2919. doi: 10.1177/0886260512438279

Nitti, M., & Ciavolino, E. (2014). A deflated indicators approach for estimating second-order reflective models through

PLS-PM: an empirical illustration. *Journal of Applied Statistics*. Volume 41, Issue 10, 2222 - 2239.

Rönkkö, M., & Cho, E. (2020). An Updated Guideline for Assessing Discriminant Validity. *Organizational Research Methods*, 1094428120968614.

Rönkkö, M., McIntosh, C. N., & Antonakis, J. (2015). On the adoption of partial least squares in psychological research: Caveat emptor. *Personality and Individual Differences*, *87*, 76-84.

Sanchez, G. (2013). *PLS path modeling with R*. Trowchez Editions, 383.

Simonsen, S., & Euler, S. (2019). Avoidant and narcissistic personality disorders. In Bateman, A., & Fonagy, P., (2019). *Handbook of Mentalizing in Mental Health Practice*. Am. Psychiatr. Publ. 2nd ed.

Tenenhaus, M. (2008). Component-based structural equation modelling. *Total Quality Management*, *19*(7–8), 871–886. doi: 10.1080/14783360802159543

Tenenhaus, M., Vinzi, V. E., Chatelin, Y. M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, *48*(1), 159–205. doi: 10.1016/j.csda.2004.03.005

Tolan, P. H., Dodge, K., & Rutter, M. (2013). Tracking the multiple pathways of parent and family influence on disruptive behavior disorders. In P. H. Tolan & B. L. Leventhal (Eds.), *Disruptive behavior disorders* (pp. 161–191). Springer.

Velotti, P., Garofalo, C., Cellea, A., Bucks, R. S., Roberton, T., & Daffern, M. (2017). Exploring anger among offenders: The role of emotion dysregulation and alexithymia. *Psychiatry, Psychology and Law, 24*(1), 128–138. doi: 10.1080/13218719.2016.1164639

Velotti, P., Garofalo, C., Petrocchi, C., Cavallo, F., Popolo, R., & Dimaggio, G. (2016a). Alexithymia, emotion dysregulation, impulsivity and aggression: A multiple mediation model. *Psychiatry Research, 30*, 237, 296-303. doi: 10.1016/j.psychres.2016.01.025

Velotti, P., Garofalo, C., D'Aguanno, M., Popolo, R., Salvatore, G. & Dimaggio, G. (2016b). Mindfulness moderates the relationship between aggression and Antisocial Personality Disorder traits: Preliminary investigation with an offender sample. *Comprehensive Psychiatry*, *64*, 38–45. https://doi.org/10.1016/j.comppsych.2015.08.004

Velotti P., Garofalo, C., Dimaggio, G., Fonagy, P. (2019). Mindfulness, Alexithymia, and Empathy Moderate Relations Between Trait Aggression and Antisocial Personality Disorder Traits. *Mindfulness, 10*(6), 1082-1090. https://doi.org/10.1007/s12671-018-1048-3

Widiger, T. A., & Trull, T. J. (2007). Plate tectonics in the classification of personality disorder: shifting to a dimensional model. *The American Psychologist*, *62*(2), 71-83.

Widiger, T. A., & Simonsen, E. (2005). Alternative dimensional models of personality disorder: finding a common ground. *Journal of Personality Disorders*, *19*(2), 110-130.

Wright, A. G. C., Pincus, A. L., Thomas, K. M., Hopwood, C. J., Markon, K. E., & Krueger, R. F. (2013). Conceptions of narcissism and the DSM-5 pathological personality traits. *Assessment*, *20*(3), 339–352. doi: 10.1177/107319111348669