

Profiles of mentalizing in individuals with antisocial behavior: Comparing state- and trait-mentalizing

| | |
|------------------|---|
| Authors | De Wit-De Visser,Brenda; Rijckmans,Madeleine J.N.; Vermunt,Jeroen K.; Hamakers,Matthijs J.W. et al |
| Published in | Psychotherapy Research |
| DOI | 10.1080/10503307.2025.2517560 |
| Publication Date | 2025-07 |
| Document Version | publishersversion |
| Link | https://research.tilburguniversity.edu/en/publications/0ed2dfffa-08d2-429f-8085-81964238d5f8 |
| Citation | De Wit-De Visser, B, Rijckmans, M J N, Vermunt, J K, Hamakers, M J W & van Dam, A 2025, 'Profiles of mentalizing in individuals with antisocial behavior : Comparing state- and trait-mentalizing', Psychotherapy Research. https://doi.org/10.1080/10503307.2025.2517560 |
| Download Date | 2026-05-17 12:23:58 |
| Rights | <p>General rights</p> <p>Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.</p> <ul style="list-style-type: none"> - Users may download and print one copy of any publication from the public portal for the purpose of private study or research. - You may not further distribute the material or use it for any profit-making activity or commercial gain - You may freely distribute the URL identifying the publication in the public portal" <p>Take down policy</p> <p>If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.</p> |

Profiles of mentalizing in individuals with antisocial behaviour: comparing state- and trait-mentalizing

| | |
|------------------|---|
| Authors | De Wit-De Visser, Brenda; Rijckmans, Madeleine J.N.; Vermunt, Jeroen K.; Hamakers, Matthijs J.W. et al |
| Published in | Psychotherapy Research |
| DOI | 10.1080/10503307.2025.2517560 |
| Publication Date | 2025 |
| Document Version | publishersversion |
| Link | https://research.tilburguniversity.edu/en/publications/8bfeb4fc-72b7-4e0b-b902-94ff3e1670a4 |
| Citation | De Wit-De Visser, B, Rijckmans, M J N, Vermunt, J K, Hamakers, M J W & van Dam, A 2025, 'Profiles of mentalizing in individuals with antisocial behaviour : comparing state- and trait-mentalizing', Psychotherapy Research. https://doi.org/10.1080/10503307.2025.2517560 |
| Download Date | 2026-02-25 08:19:02 |
| Rights | <p>General rights</p> <p>Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.</p> <ul style="list-style-type: none"> - Users may download and print one copy of any publication from the public portal for the purpose of private study or research. - You may not further distribute the material or use it for any profit-making activity or commercial gain - You may freely distribute the URL identifying the publication in the public portal" <p>Take down policy</p> <p>If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.</p> |



Profiles of mentalizing in individuals with antisocial behavior: comparing state- and trait-mentalizing

Brenda De Wit-De Visser, Madeleine J.N. Rijckmans, Jeroen K. Vermunt, Matthijs J.W. Hamakers & Arno van Dam

To cite this article: Brenda De Wit-De Visser, Madeleine J.N. Rijckmans, Jeroen K. Vermunt, Matthijs J.W. Hamakers & Arno van Dam (03 Jul 2025): Profiles of mentalizing in individuals with antisocial behavior: comparing state- and trait-mentalizing, *Psychotherapy Research*, DOI: [10.1080/10503307.2025.2517560](https://doi.org/10.1080/10503307.2025.2517560)

To link to this article: <https://doi.org/10.1080/10503307.2025.2517560>



© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 03 Jul 2025.



Submit your article to this journal [↗](#)



Article views: 1294



View related articles [↗](#)



View Crossmark data [↗](#)

RESEARCH ARTICLE

Profiles of mentalizing in individuals with antisocial behavior: comparing state- and trait-mentalizing

BRENDA DE WIT-DE VISSER ^{1,3}, MADELEINE J.N. RIJCKMANS ^{2,5}, JEROEN K. VERMUNT ⁴, MATTHIJS J.W. HAMAKERS ^{1,3}, & ARNO VAN DAM ^{1,3}

¹GGZ WNB, Research and Innovation, Halsteren, The Netherlands; ²Fivoor, Fivoor Science & Treatment Innovation, Poortugaal, The Netherlands; ³Tilburg School of Social and Behavioral Sciences, Tranzo Scientific Center for Care and Welfare, Tilburg University, Tilburg, The Netherlands; ⁴Tilburg School of Social and Behavioral Sciences, Department of Methodology and Statistics, Tilburg University, Tilburg, The Netherlands & ⁵Tilburg School of Social and Behavioral Sciences, Department of Developmental Psychology; Clinical and Forensic Psychology, Tilburg University, Tilburg, The Netherlands

(Received 4 December 2024; revised 31 May 2025; accepted 2 June 2025)

Abstract

Objectives: Mentalizing is a crucial factor in understanding antisocial behavior. The current study focuses on mapping mentalization within a population of patients with antisocial behavior ($n = 108$) and compares trait and state mentalizing.

Methods: Three instruments were used to assess mentalizing abilities: the Reflective Functioning Scale, an emotion recognition task, and a Virtual Reality experiment. Mentalizing profiles were determined with latent class analyses and subsequently examined their relations.

Results: Patients with antisocial behavior exhibited poor mentalizing capacities. They showed problems with general mentalizing capacities, specific problems in emotion recognition and reduced emotional reactivity. Half of the participants displayed reduced trust towards others during state mentalizing, indicating imbalances between automatic and controlled mentalizing. A part of the population showed hostile attribution bias, related to increased anger and threat perception and reduced experienced trust in direct social interaction. State-mentalizing and trait-mentalizing were not related.

Conclusion: The findings align with previous studies on mentalizing in individuals with antisocial personality disorder. However, this study underscores the importance of further investigating the heterogeneity of mentalizing capacity within this population, especially in comparing state- and trait mentalizing. Mapping underlying mentalizing patterns of these patients may provide directions for tailoring therapeutic interventions.

Clinical or methodological significance of this article: Addressing mentalizing problems is crucial in treating antisocial behavior, yet little research exists on mentalizing profiles and state vs. trait mentalization. This study highlights variability in mentalizing capacities, including impairments in emotion recognition and social interaction. Understanding these patterns helps tailor interventions to specific challenges.

Introduction

Mentalizing capacity has developed into an important psychotherapeutic concept over the past decade. It receives increasing interest due to its

perceived importance in the development of various mental disorders (Bateman et al., 2019; Bateman & Fonagy, 2016). Mentalization refers to the ability to perceive or interpret one's own and others' mental

Correspondence concerning this article should be addressed to Brenda de Wit-de Visser, Email: a.c.devisser@tilburguniversity.edu GGZ Westelijk Noord-Brabant, Hoofdlaan 8, Halsteren 4661 AA, The Netherlands

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

states in relation to behavior (Bateman et al., 2013). By mentalizing about oneself and others, individuals make an image of intentions, motivations, and emotions. In literature, several concepts are closely related to mentalization, including theory of mind, emotion recognition, empathic capacities, and affective resonance. Mentalization can thus be viewed as a multidimensional construct, wherein several dimensions may be present to varying degrees in each individual (Bateman & Fonagy, 2016).

One of the psychological phenomena where mentalization seems to be important is antisocial behavior (Abi-Habib et al., 2020; Bateman et al., 2013; Newbury-Helps et al., 2017). Antisocial behavior is a transdiagnostic construct (De Wit-De Visser et al., 2023), which is primarily linked to the Antisocial Personality Disorder (ASPD), but is also associated with many other disorders, such as Intermittent Explosive Disorder (IED), Conduct Disorder (CD), ADHD, addiction problems, or other cluster-B personality disorders and psychopathy (Kotov et al., 2017; Rijckmans et al., 2020), and can coexist with various mental disorders. For example, individuals known to have ADHD, addiction problems, or cluster-B personality problems may exhibit antisocial behavioral patterns (Rijckmans et al., 2020).

Well-developed mentalizing capacity can serve as an inhibiting factor for antisocial behavior. However, due to traumatic experiences, congenital vulnerabilities (such as developmental disorders), or difficulties in the interaction with the attachment relationship, the development of one's mentalizing ability can be delayed or inhibited (Gergely & Unoka, 2008). According to Blair (1995), the violence inhibition mechanism is a neurobiological network that inhibits violence upon perceiving distress in others. If the ability to mentalize is impaired by deficits in detecting distress cues, the threshold for engaging in violent behavior is lower (Blair, 1995). General low mentalizing capacities are found in individuals with antisocial behavior (Abi-Habib et al., 2020; Newbury-Helps et al., 2017).

On the other hand, specific problems in social information processing may also enhance vulnerability to engage in antisocial behavior. In social information processing the first step involves perceptive and sensory processes in coding social cues, and, subsequently, the second step involves causal attribution to the same cues in terms of intentional attribution (Dodge & Crick, 1990). A broadly reported underlying mechanism of antisocial behavior is hypervigilance for hostility or threat (De Castro et al., 2002; Klein Tuente et al., 2019), which occurs in these first steps of social information processing.

Studies have shown significant associations between emotional abuse and neglect and (mis)trust,

suggesting that childhood traumatic experiences shape social information processing, potentially leading to increased mistrust (Benzi et al., 2023). Milesi et al. (2023) outlined a theoretical framework in which epistemic trust – a facet of interpersonal trust – and facial trustworthiness are elements of a more complex developmental model of trust. Interpersonal trust develops during interactions in a secure attachment. Facial expressions (ostensive cues) contribute to the development of epistemic trust, which is the “openness to the reception of social communication that is personally relevant and of generalizable significance” (Bateman et al., 2019). Early trustworthiness judgements in a secure childhood contribute to the development of epistemic and interpersonal trust (Milesi et al., 2023). However, prior negative trust experiences can lead to epistemic mistrust and may contribute to fostering a biased perception of social reality (Bateman & Fonagy, 2012).

Epistemic mistrust can lead to the misinterpretation of social signals in daily life (Bateman et al., 2019). Trust attribution occurs in two stages: an initial, implicit evaluation of ostensive cues, followed by a more profound judgment involving interpersonal and epistemic trust (Milesi et al., 2023). However, when social signals are misinterpreted, for instance in a situation of perceived threat (perception of imminent physical danger or an untrustworthy opponent), mentalizing may be temporarily impaired, which may escalate into aggressive behavior when someone feels the need to protect oneself. These issues in social interaction present themselves at the level of momentary mentalizing, commonly referred to as state-mentalizing. From this, it can be inferred that both general difficulties in mentalizing (at the level of personality traits; trait mentalizing) and specific difficulties in mentalizing, such as at the level of state mentalizing, may contribute to the development of antisocial behavioral problems.

The understanding of underlying mentalizing profiles in other disorders, such as borderline personality disorder, is already well-developed (Bateman & Fonagy, 2012). In this context, specific imbalances across various dimensions of mentalizing are clearly identified. Insight into mentalizing deficits is crucial for better understanding the problems associated with these disorders and for tailoring interventions to address these underlying problems. However, when it comes to antisocial behavior, there is still much ignorance of the underlying mentalizing problems and mentalizing capacities appear to vary within the population.

Bateman and Fonagy (2012) described a mentalizing profile for patients with ASPD. They distinguish four dimensions of mentalizing; automatic-

controlled-, self-other-, external-internal-, and cognitive-affective mentalizing. In mentalizing, a balance between these dimensions is crucial. Bateman and Fonagy (2012) state that patients with ASPD tend to overuse automatic mentalizing, lacking sufficient reflection on their assumptions about themselves and others through controlled mentalizing. Problems with resonating with others' emotions and reduced interest in others explain their limited ability for controlled mentalizing. They also linked ASPD to a cognitive way of mentalizing and a lack of affective connection with oneself and others. Patients with ASPD can reason about mental states in a cognitive way, without truly understanding feelings of oneself or others. A possible explanation for problems in affective mentalizing may be hypoactivation of the amygdala (Bateman et al., 2013), leading to reduced responsiveness to distress cues. This complicates empathizing with others and may result in self-serving behavior. Subsequently, Bateman and Fonagy (2012) describe that patients with ASPD either misuse their capacity to read others' minds (cognitive mentalizing) for their own purposes or excessively focus on their own needs and how they can use others to fulfill them. This results in a fixation on either the self or the other pole of the self-other dimension, both leading to one-sided relationships. The focus on the self is also reflected in the internal-external dimension of mentalizing. Individuals with ASPD are primarily focused on their own internal world and needs, lacking the ability to reflect on it. The ability to recognize others' emotions is also limited. Problems with reading others' emotions, or misreading them, lead to difficulties in taking others' perspective and therefore problems in social interactions. By presenting a mentalizing profile for patients with ASPD, Bateman and Fonagy (2012) provided important directions for therapeutic interventions. However, insufficient empirical research has been conducted on these various dimensions of mentalizing in patients with antisocial problems.

The presented framework by Bateman and Fonagy (2012) offers a foundational understanding of ASPD, however, this is hypothetical in nature. There is a demand for empirical evidence of mentalization deficits in patients with antisocial behavior, to determine whether these deficits are uniform across the population. Gagliardini et al. (2023) examined mentalizing profiles in individuals with personality pathology and identified four mentalizing clusters; the Affective-Self-Automatic (ASA) Profile, the Other-Automatic-Affective (OAA) Profile, the External (E,) Profile, and the Cognitive-Self-Automatic (CSA) Profile. They discovered that individuals with ASPD were predominantly represented in

the ASA-profile (76%), indicating a focus on the self, imbalanced affective mentalizing, and reliance on automatic mentalizing. This pattern was also observed in other cluster-B personality disorders (Narcissistic and Borderline personality disorder). A smaller portion of individuals with ASPD was categorized under the OAA-profile (14%), which similarly exhibited a focus on automatic mentalizing and imbalance in affective mentalizing, but with a shift towards focusing on others rather than themselves compared to the ASA-profile. In contrast, the E- and CSA-profile were less prevalent (5% each). These results demonstrate that disbalances in mentalizing abilities are significant shared factors within personality psychopathology and transcend the boundaries of specific disorders. Furthermore, it underlines that various mentalizing profiles may underlie psychopathology. This underscores that a comprehensive understanding of mentalizing profiles can contribute to a deeper comprehension of broad antisocial psychopathology.

Fonagy and Luyten (2018) followed the Research Domain Criteria (RDoC) approach and presented "mentalizing disfunctions" as a basis for conduct problems in youth, including issues with perspective-taking and empathy. Importantly, they described youth with conduct problems as a heterogeneous population, where problems in mentalizing are expressed in various ways and also may have different underlying causes. Fonagy and Luyten (2018) differentiate between youth with high and low callous-unemotional (CU) traits. Youth with low empathic capacities and a callous approach of others (high CU-traits) experience low levels of anxiety and are therefore less responsive to threat. In contrast, youth with low CU-traits are more susceptible for threat and stress, experiencing high levels of anxiety. This could form the basis for the reported differences in mentalization in literature. Individuals with high CU traits seem to have problems with recognizing distress cues (fear, sadness, pain) (Blair et al., 2001; Fonagy & Luyten, 2018), whereas those with low CU traits are likely more hypervigilant to threat (Fonagy & Luyten, 2018). Fonagy and Luyten (2018) described the mentalizing profile of individuals with high CU-traits as excessively cognitively oriented and impaired in affective mentalizing. On the other hand, individuals with low CU-traits predominantly engage in automatic mentalizing modes due to their hypervigilance to threat. Although both subgroups experience problems with interpersonal understanding, there are clear differences in their mentalizing profiles.

Recently, De Wit-De Visser et al. (2023) introduced a conceptual model that elucidates various underlying dynamics of antisocial pathology. Within

this model, mentalizing is portrayed as a crucial factor in the development and perpetuation of antisocial behavior. In line with Fonagy and Luyten (2018), they present various mentalizing profiles that may underlie antisocial behavior. A bias towards hostility and hypersensitivity for threat, reduced cognitive and affective mentalizing, and specific deficits in affective mentalization are presented as explanatory factors in antisocial problems. They considered that there may be general deficits in mentalizing (e.g., hypomentalizing) as well as deficits that emerge in specific contexts (e.g., hostile bias in direct social interaction). This approach emphasizes the heterogeneity within the antisocial population regarding underlying explanations for antisocial behavior.

Bateman and Fonagy (2012) laid the groundwork for a deeper comprehension of the role of mentalizing functioning in individuals with ASPD by delineating a mentalizing profile. However, more recently, questions have been raised regarding the homogeneity of this population and whether there is a consistent profile of mentalizing abilities. Fonagy and Luyten (2018), Gagliardini et al. (2023), and De Wit-De Visser et al. (2023) suggest heterogeneity in mentalizing abilities and have contributed to an enhanced understanding of diverse mentalizing profiles in the development of antisocial psychopathology. Given the hypothetical nature of most studies and the limited availability of data on individuals with antisocial problems, further empirical research on these theories is needed to enhance understanding of mentalizing functioning in individuals struggling with antisocial behaviors. Furthermore, mentalizing functioning can vary depending on the context. While trait mentalizing describes a more stable ability of mentalizing, state mentalizing refers to the ability to mentalize in a specific context (such as in direct interaction with others) (Luyten et al., 2024). Generally, individuals may possess an adequate capacity for mentalizing; however, this ability can temporarily diminish under high arousal. There is insufficient research available on the distinction and interrelationship between trait- and state mentalizing in antisocial populations. Therefore, it is crucial to further investigate how mentalizing abilities manifest within this population.

The primary objective of this study is to investigate whether there is a specific mentalizing profile in a sample of patients with antisocial behavioral problems. Knowledge about mentalizing functioning may guide diagnostics and treatment of individuals with antisocial behavior. Varying mentalizing dysfunctions may require distinct therapeutic approaches. To investigate mentalizing functioning, this study utilized a sample of individuals exhibiting antisocial behaviors for more than one year, as antisocial behavior can

occur across various psychological disorders and can therefore be considered a transdiagnostic phenomenon (De Wit-De Visser et al., 2023).

The first research question examines whether deficits in mentalization are observable within a population of individuals with antisocial behavior. The second research question is whether there are different profiles of mentalizing in this population. Based on the studies of Fonagy and Luyten (2018), Gagliardini et al. (2023), and De Wit-De Visser et al. (2023) our hypothesis is that various mentalizing deficits (subgroups) can be observed, including hostile attribution bias (heightened sensitivity to anger and threat), reduced overall mentalizing ability, and specific impairments in affective mentalizing. This study will investigate reflective functioning (trait-mentalizing) as well as the capacity for emotion recognition and the perception and interpretation of ostensive cues (anger, threat, and trustworthiness) in direct social interaction as important components of mentalizing (Bateman et al., 2013). Given that mentalizing capacity can change under increasing arousal, a virtual simulation environment will be used to assess mentalizing in direct social interaction (i.e., state-mentalizing). We expect, based on the studies of Fonagy and Luyten (2018) and De Wit-De Visser et al. (2023), that under certain circumstances (e.g., high arousal), state-mentalizing is not related to trait-mentalizing. Individuals may maintain stability in mentalizing under regular conditions but lose this capacity under high arousal (De Wit-De Visser et al., 2023; Fonagy & Luyten, 2018; Nolte et al., 2013).

Methods

Participants

For the current study, 108 participants were recruited in Mental Health Center West North Brabant (MHC WNB) and Fivoor forensic and intensive psychiatry (Fivoor) in the Netherlands. Recruitment took place in general and specialized mental health settings (including inpatient and outpatient settings) and forensic settings (including forensic inpatient and outpatient settings, and patients with detention under hospital order (Dutch: TBS)). Patients were eligible for participations if they were aged between 18 and 65 years old and have a mental illness (diagnosis of a psychiatric disorder). They needed to be known with at least one form of antisocial behavior.

Antisocial behavior is conceptualized as Physical and Verbal Aggression, Social Aggression and Rule-breaking behavior (three facets of antisocial behavior as found by Burt and Donnellan, 2009). These forms can be reactive as well as proactive (Raine

et al., 2006). Antisocial behavior will be screened with a Dutch version of the Externalizing Spectrum Inventory (ESI-bf; (Soe-Agnie et al., 2015)) and de Reactive Proactive Questionnaire (RPQ; (Cima et al., 2013)). The subscales Physical Aggression, Relational Aggression, Theft, Fraud, Rebelliousness, and Destructive Aggression of the ESI were strongly correlated with the three facets of antisocial behavior as mentioned by Burt and Donnellan (2009), and therefore used to screen for antisocial behavior together with the RPQ. Patients needed to recognize minimally one form of antisocial behavior on one of these subscales of the ESI or RPQ. Furthermore, antisocial behavior is to be expressed for a minimum of 1 year and antisocial behavior should not only be presented in acute episodes of psychiatric disorders (e.g., acute episodes of psychosis or use of narcotics). Patients were excluded when they experienced acute symptoms of suicidality at the moment of participation. Also, illiteracy or not mastering the Dutch language are exclusion criteria for participation, since data collection requires the understanding of the Dutch language. Patients who have or had the medical condition “epilepsy” could take part in data collection, however they were excluded from testing with Virtual Reality.

Procedure

The current study was approved by the Medical Ethical Committee (METC) as part of a larger study (NL76121.028.21), involving research which is subject to the Medical Research Involving Human Subjects Act (WMO). The study protocol was also endorsed by the mental health centers where data were collected (MHC WNB and Fivoor). The study was conducted according to the principles of the Declaration of Helsinki (Version 64th WMA General Assembly, Fortaleza, Brazil, October 2013) and in accordance with the Medical Research Involving Human Subjects Act (WMO). This study was not preregistered; however, the hypotheses were based on the study by De Wit- de Visser and colleagues (2023), as described in the Introduction. For participation, patients received a 40-euro incentive. The used instruments were administered as part of a broader data collection process consisting of two test sessions. The AAI/RFS was administered during the first test session. The ERART and VR-AC were administered during the second test session, in the order described.

Measures

Screening for antisocial behavior. Externalizing Spectrum Inventory, brief form (ESI-

bf). The ESI-bf was used to measure antisocial problems. The ESI-bf contains 160 items about broad antisocial problems during lifetime. Items are scored on a 4-point scale (true, somewhat true, somewhat false, false). The Dutch version was used for the current study (Soe-Agnie et al., 2015). The reliability (i.e., Cronbach’s alpha) for the used subscales is: Physical Aggression (.88), Relational Aggression (.89), Theft (.90), Fraud (.84), Rebelliousness (.91), and Destructive Aggression (.92).

Reactive-Proactive Aggression Questionnaire (RPQ). The RPQ was used to measure reactive and proactive aggression in the antisocial population. The validated Dutch version was included for this study (Cima et al., 2013). The RPQ consists of 23 items. 11 items measure reactive aggression, and 13 items measure proactive aggression. Questions are answered on a 3-point scale (0 = never, 1 = sometimes, 2 = often), where participants score on how often they recognize a specific statement (example: “Gotten angry when others threatened you”). The internal reliability (Cronbach’s alpha) in the current study for the reactive aggression subscale is .87 and for the proactive aggression subscale .90.

Mentalizing capacities. Adult Attachment Interview (AAI) and Reflective Functioning Scale (RFS). The Dutch translation of the AAI (Bakermans-Kranenburg & van IJzendoorn, 1993) was administered to participants to measure trait-mentalizing with the Reflective Functioning Scale (RFS (Fonagy et al., 2016)). The AAI consists of an interview (+/–1 h) with questions about attachment in childhood. The AAI was transcribed and scored with the RFS to measure the level of mentalizing capacity. A score from –1 (antireflective, hypomentalizing) to 9 (exceptionally reflective) was used as measure for mentalizing functioning. Only trained clinicians can score the RFS (3 days training). Coders are science-practitioners (psychologist/mental health-psychologist and researcher) and trained by the Anna Freud Institute. In the current study, 11 interviews were double coded to assess interrater reliability. The intraclass correlation of the RFS-score was excellent (ICC = .93). The RFS is the golden standard for the measurement of reflective functioning.

Emotion Recognition and Affective Resonance Task (ERART). The ERART (developed by coordinating researcher) is a computerized task for the measurement of emotion recognition and affective resonance (affective mentalizing). The ERART is based on the Emotion Recognition Task (ERT) developed by Montagne et al. (2007). The test is built up in OpenSesame, an experiment-builder for

social sciences (Mathot et al., 2012). The test starts with 4 questions following the Outcome Rating Scale (Miller & Duncan, 2000) to measure participants' wellbeing at the moment of testing. Subsequently, two examples are presented to make participants familiar with the test. The test contains 54 screens with a stimulus (short videoclips with frontal view of Caucasian individuals) of a morphed emotion (intensity of 0% (neutral), 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%). Since Montagne et al. (2007) found that emotions below 20% are difficult to recognize correctly, the current study starts with morphed emotions at 30% emotion intensity. The stimuli are interactive morphing faces (videoclips) presenting emotions from neutral to a specific emotion. The following emotions are presented: angry, happy, anxious, disgusted, sadness, surprised and neutral. Each stimulus is accompanied with a question for emotion recognition ("Which emotion do you see?") and affective resonance ("What do you feel when others in your environment feel like this?"). Participants can answer both questions by multiple-choice (angry, happy, anxious, disgusted, sadness, surprised and neutral) using a forced-choice method (without time-limit). The emotion can be selected via mouse click. More information about the stimuli can be found in the study of Montagne et al. (2007). Since the data from the ERART are used to define the latent variables, reliability is assessed using the entropy R^2 score. The reliability of the individual subscales is of less relevance in this context, as the primary focus is on how well the total scores of the ERART reflect the latent variables and predict the latent classes (for a thorough discussion on why

Cronbach's alpha is not the best predictor of reliability for emotion recognition tasks, see: Williams et al. (2023)). As shown in the results (Table V), the latent classes demonstrate a (very) high predictive value (entropy $R^2 = .85$ for a 3-class model).

Aggression Catwalk in Virtual Reality (AC-VR). The aggression catwalk (AC) (produced by CLEVR B.V.; <https://clevr.net>) in Virtual Reality (VR) measures the amount in which persons perceive others as angry, threatening and (dis)trustworthy. The AC-VR consists of an interaction with a virtual character whom displays ostensive cues with a specific level of verbal aggression. Level 1 represents no aggression (friendly conversation), Level 2 indicates mild irritation, Level 3 shows increasing aggression/anger, and Level 4 involves strong verbal and non-verbal aggression. This was presented by facial emotions, body language and verbal aggression. A participant was asked to score the level of anger of the virtual character by answering the following question: "How angry does this person seem to you?" on a scale from 0 to 10 for (0 not angry to 10 very angry), the level of experienced threat ("How threatening does this person seem to you?"; 0 not threatening to 10 very threatening) and the level of trustworthiness of the virtual character ("To what extent do you find this person trustworthy?"; 0 not trustworthy to 10 very trustworthy). At the start of the VR-task, two practice scenarios are presented to become familiar with the task. After that, 16 series of test-interactions were presented (4 per level of aggression), each involving the above mentioned three questions (anger, threat, and trustworthiness). A picture of a virtual test scenario can be found in Figure 1. The AC-VR was used to assess how



Figure 1. Example virtual test scenario (AC-VR).

individuals perceive and interpret others behavior and mental states in an interactive context, which is indicative for state-mentalizing (mentalizing in the moment). Specific, the AC-VR indicates whether some form of mentalizing has occurred regarding external information (external mentalizing), and whether individuals can perceive and interpret the mental states of others on affective level (affective mentalizing). The varying level of aggression of the virtual characters, and associated varying levels of arousal, make it possible to reflect on the dimension of automatic mentalizing. Furthermore, the estimated level of trustworthiness in the neutral condition is indicative of underlying (interpersonal) trust, because there are no clues to substantiate a realistic judgement. Since the data from the AC-VR are used to define the latent variables, reliability is assessed using the entropy R^2 score. The reliability of the individual subscales is of less relevance in this context, as the primary focus is on how well the total scores of the AC-VR reflect the latent variables and predict the latent classes. As shown in the results (Table VII), the latent classes demonstrate a (very) high predictive value (entropy $R^2 = .94$ for a 5-class model).

Data Analysis

To explore the mentalizing capacities of individuals with antisocial problems and differentiate between mentalizing profiles, latent class analysis (LCA) was performed for both the data collected within the ERART and the AC-VR task using LatentGOLD 6.0 (Vermunt & Magidson, 2021). LCA is a probabilistic clustering method that allows for the identification of subgroups characterized by specific response patterns on observed variables. The ERART consisted of 8 tasks for each of the 6 emotions anger, sadness, surprise, fear, disgust, and happiness. The number of correct emotion recognitions out of 8 tasks were treated as 6 binominal count class indicators. The responses on the VR tasks were combined (by computing the average) per degree of aggression, and subsequently coded into 5 ordinal categories. This yielded four ordinal indicators for assessed anger, threat, and trust, thus 12 class indicators in total. To avoid local maxima, all models were estimated using 160 start sets and 250 initial iterations per start set. Class enumeration was mainly based on the Bayesian information criterion (BIC). However, we also looked at the Akaike information criterion (AIC), AIC3, the maximum bivariate residual (BVR), and the Vuong-Lo-Mendel-Rubin (VLMR) test. After deciding on the number of classes, patients were assigned to the latent classes based on their posterior class

membership probabilities. Next, the association between the ERART and VR class memberships was explored, as well as the association between these two class memberships and mentalizing capacities measured by the RFS. For this purpose, we used the biased adjusted step-three approach implemented in LatentGOLD (Bakk & Vermunt, 2016; Bolck et al., 2004; Vermunt, 2010).

Results

Descriptives of the samples used in the current study are presented in Table I.

To gain insight into the general mentalizing abilities of individuals with antisocial behavior, descriptive data are presented on the distribution of trait mentalizing (RFS) and emotion recognition across the entire sample.

Descriptives Reflective Functioning Scale

The descriptive statistics of the RFS based on the AAI-interview can be found in Table II. Figure 2 demonstrates the distribution of RFS-scores in the current sample.

Descriptives Emotion Recognition (ERART)

Table III presents the descriptives of emotion recognition on the ERART. False positives of anger are reported in Table IV. These false positives are used as an indicator of hostile attribution bias.

Latent Class Analysis – Emotion Recognition (ERART, $n = 103$)

Latent class analyses were conducted to gain insight into differences within the population regarding mentalizing. The aim is to determine whether specific mentalizing deficits can be identified within the latent classes (subgroups).

Table V presents the statistics for the one to six class model estimated with the ERART data. The BIC selected a model with 3 classes, the AIC3 with 4 classes, and the AIC with 6 or more classes. It can also be noted that the maximum BVR dropped firmly till 3 classes, but not much anymore after this. Based on the BIC, we decided to retain the model with 3 classes. This model has an entropy R^2 value of .85 showing that class separation is very good.

Table VI reports the overall means and the class-specific means of the six class indicators, as well as the overall Wald test for the differences between the three classes and the pairs of classes for which the pairwise Wald test is significant (based on $p <$

Table I. Sociodemographic characteristics and antisocial behaviour.

| Variables | RFS Subset | | ERART Subset | | VR Subset | |
|---|------------|---------------|--------------|---------------|-----------|---------------|
| | (valid) N | Mean (SD) % | (valid) N | Mean (SD) % | (valid) N | Mean (SD) % |
| Age | 78 | 43.33 (12.54) | 108 | 42.48 (12.43) | 103 | 42.57 (12.46) |
| Gender | 78 | | 108 | | 103 | |
| Male | 58 | 74.40% | 80 | 74.10% | 76 | 73.80% |
| Female | 19 | 24.40% | 27 | 25.00% | 26 | 25.20% |
| Other | 1 | 1.30% | 1 | 0.90% | 1 | 1% |
| Highest educational level | 78 | | 108 | | 103 | |
| No education | 3 | 3.80% | 5 | 4.60% | 4 | 3.90% |
| Primary school | 9 | 11.50% | 14 | 13.00% | 14 | 13.60% |
| Pre-vocational education (lbo/vmbo/mavo) | 20 | 25.70% | 27 | 25.00% | 25 | 24.30% |
| Senior general secondary education (havo) | 4 | 5.10% | 4 | 3.70% | 4 | 3.90% |
| University preparatory education (vwo) | 3 | 3.80% | 6 | 5.60% | 6 | 5.80% |
| Vocational education (mbo) | 23 | 29.50% | 29 | 26.90% | 30 | 29.10% |
| Higher vocational education (HBO) | 9 | 11.50% | 14 | 13.00% | 12 | 11.70% |
| University (WO) | 3 | 3.80% | 3 | 2.80% | 3 | 2.90% |
| Other | 4 | 3.80% | 6 | 5.80% | 5 | 4.90% |
| Relational Aggression (ESI) | 77 | 9.14 (6.43) | 105 | 8.64 (6.25) | 100 | 8.65 (6.21) |
| Physical Aggression (ESI) | 77 | 10.91 (7.21) | 105 | 10.38 (6.97) | 100 | 10.36 (6.86) |
| Rulebreaking Behavior (ESI) | 77 | 23.55 (18.16) | 105 | 24.20 (19.29) | 100 | 23.83 (18.98) |
| Proactive Aggression (RPQ) | 77 | 5.68 (4.73) | 105 | 5.44 (4.76) | 100 | 5.36 (4.46) |
| Reactive Aggression (RPQ) | 77 | 12.31 (4.87) | 105 | 12.14 (4.88) | 100 | 12.19 (4.87) |

Note: In the descriptive data of the RFS, ESI and RPQ, the data of one participant was excluded due to missing data which may have affected the descriptives. Two participants were included based on their residence in clinical forensic setting and history of persistent antisocial behaviour.

Table II. Descriptive statistics RFS.

| | N | Min | Max | Mean | SD |
|-----------|----|-----|-----|------|------|
| RFS-score | 78 | -1 | 6 | 2.97 | 1.96 |

0.05). The class-specific means for the three emotion recognition profiles are also depicted in Figure 3. Class 1 comprised 50,5% of the sample ($n = 52$) and represents patients with better recognitions of all emotions compared to the other two classes. We will refer to this class as the high emotion recognition

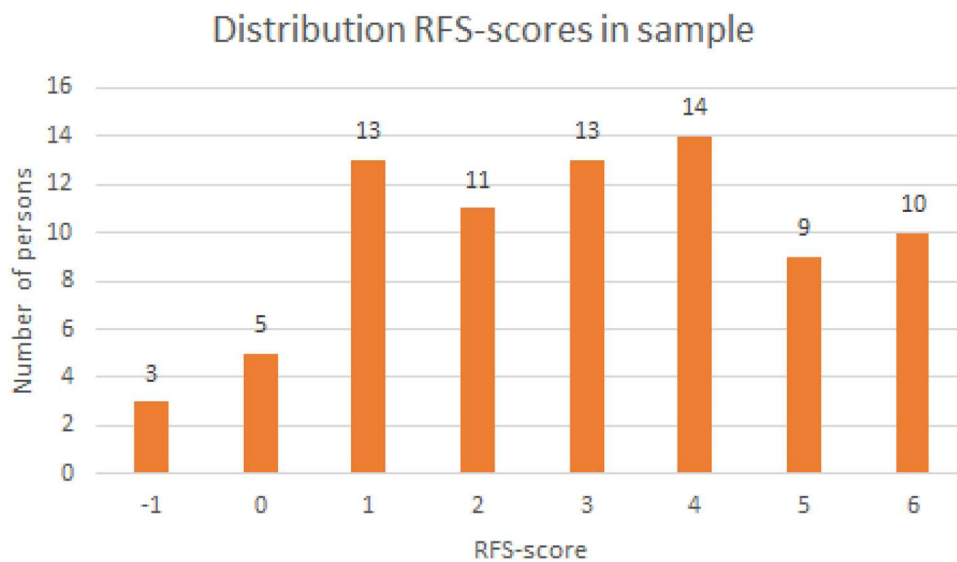


Figure 2. Distribution of RFS-scores in sample.

Table III. Descriptives emotion recognition ERART.

| | <i>N</i> | <i>Min</i> | <i>Max (=8)</i> | <i>Mean</i> | <i>SD</i> |
|-----------|----------|------------|-----------------|-------------|-----------|
| Anger | 103 | 0 | 8 | 5.95 | 1.64 |
| Sadness | 103 | 0 | 6 | 1.92 | 1.40 |
| Surprise | 103 | 0 | 7 | 4.59 | 1.65 |
| Disgust | 103 | 0 | 8 | 4.40 | 2.38 |
| Happiness | 103 | 0 | 8 | 6.56 | 1.42 |
| Fear | 103 | 0 | 6 | 1.50 | 1.65 |

Table IV. Descriptives false positives of anger (hostile attribution bias).

| | <i>N</i> | <i>Min</i> | <i>Max</i> | <i>Mean</i> | <i>SD</i> |
|------------|----------|------------|------------|-------------|-----------|
| Anger bias | 103 | 0 | 11 | 2.51 | 2.42 |

profile (high ER-profile). The second class was composed of 42.1% of the sample ($n = 43$) and represents patients that experience more difficulties with recognizing emotions compared to class 1, particularly in case of negative emotions. These patients scored very low on the emotions disgust and fear compared to the high ER-profile. They also scored significantly lower on the emotions anger and surprise compared to the high ER-profile. We named this class the Impaired Negative Emotions-profile (INE-profile). The third class encompassed 7.4% of the sample

($n = 8$) and is characterized by a general low recognition of all emotions. These patients scored significantly lower on all emotions compared to patients in class 1. Compared to class 2, they scored significantly lower on the emotions anger, surprise, and happiness. They also scored lower on sadness; however, this difference is not statistically significant. We named this class the low emotion recognition profile (low ER-profile).

Latent Class Analysis – VR

Table VII presents the statistics for the one to six class models for the experienced anger and threat and assessed trust in social interaction with a Virtual Reality-character (state-mentalizing). According to the BIC, a model with 5 classes should be selected, while AIC and AIC3 statistics point at 6 or more classes. Though the maximum BVR remained somewhat large, it did not decrease markedly with an increase in the number of classes. Therefore, based on the BIC, the 5- class model was selected. The entropy R^2 indicates an excellent prediction of class membership (Table VII).

Table VIII presents the overall and class-specific means of the VR items, as well as the overall Wald tests and the significant Wald tests between pairs of classes. The means of the 5 state-mentalizing profiles measured within the AC-VR are also depicted in Figure 4. These profiles are based on the experienced

Table V. Fit indices for the latent class models estimated with the ERART data ($n = 103$).

| Number of classes | LL | BIC | AIC | AIC3 | MAX. BVR | VLMR | ENTROPY R^2 |
|-------------------|----------|---------|---------|---------|----------|-----------|---------------|
| 1 | -1214.28 | 2456.37 | 2440.56 | 2446.56 | 40.34 | | |
| 2 | -1115.05 | 2290.34 | 2256.09 | 2269.09 | 13.60 | 198.47*** | 0.84 |
| 3 | -1096.57 | 2285.83 | 2233.13 | 2253.13 | 5.90 | 36.97* | 0.85 |
| 4 | -1084.64 | 2294.43 | 2223.28 | 2250.28 | 4.65 | 23.85* | 0.77 |
| 5 | -1072.75 | 2303.08 | 2213.50 | 2247.50 | 4.20 | 23.78* | 0.78 |
| 6 | -1064.36 | 2318.74 | 2210.72 | 2251.72 | 2.18 | 16.78 | 0.80 |

Note: LL = log-likelihood; BIC = Bayesian information criteria; AIC = Akaike information criteria; AIC3 = AIC with 3 as penalized factor; Max. BVR = maximum of bivariate residual; VLMR = Voug-Lo-Mendell-Rubin.

* $p < .05$, *** $< .001$.

Table VI. Overall and class-specific means of the six ERART indicators, overall Wald tests, and significant pairwise Wald tests for the 3 ER-profiles ($n = 103$).

| | <i>Overall mean</i> | <i>Class 1</i> | <i>Class 2</i> | <i>Class 3</i> | <i>Wald</i> | <i>Paired comparisons</i> |
|-----------------|---------------------|----------------|----------------|----------------|-------------|---------------------------|
| ER of anger | 5.95 | 6.58 | 5.74 | 2.70 | 43.58*** | 3 < 1,2; 2 < 1 |
| ER of sadness | 1.92 | 2.19 | 1.78 | 0.86 | 7.21* | 3 < 1 |
| ER of surprise | 4.59 | 5.23 | 4.20 | 2.35 | 27.79*** | 2,3 < 1; 3 < 2 |
| ER of disgust | 4.40 | 6.12 | 2.57 | 2.95 | 116.51*** | 2,3 < 1 |
| ER of happiness | 6.56 | 6.87 | 6.62 | 4.04 | 25.62*** | 3 < 1,2 |
| ER of fear | 1.49 | 2.45 | 0.44 | 0.95 | 50.50*** | 2,3 < 1 |

Note: * $p < .05$, *** $< .001$; paired comparisons with $p < 0.05$.

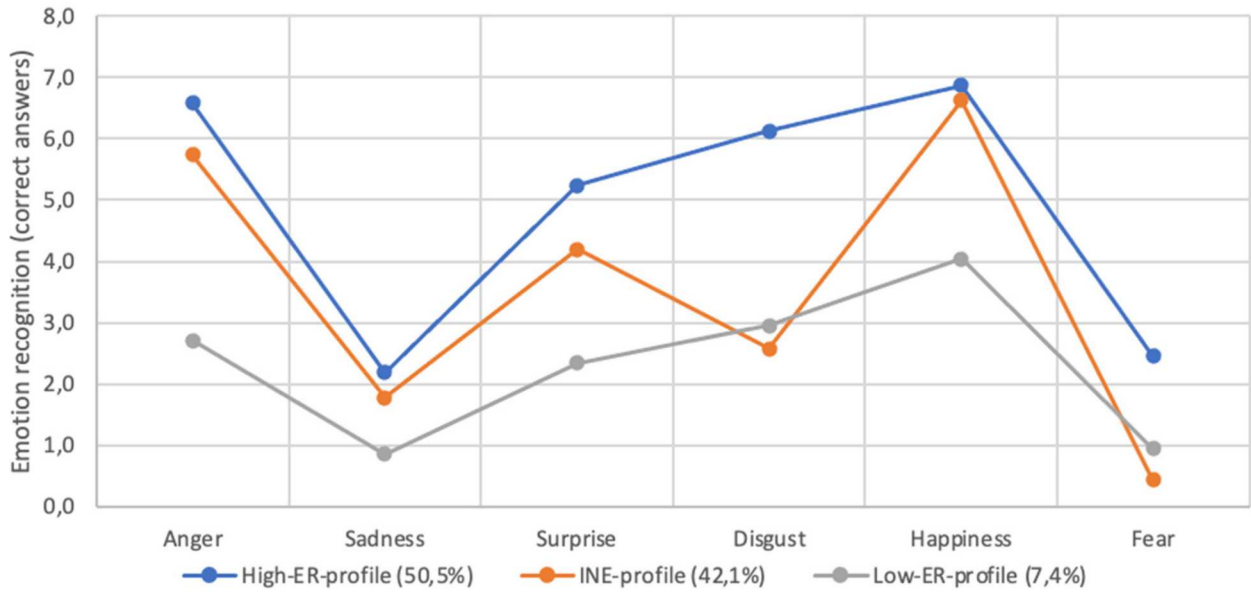


Figure 3. Emotion recognition profiles. Note: = High-ER-profile = high emotion recognition profile; INE-profile = impaired negative emotions profile; Low-ER-profile = low emotion recognition profile.

Table VII. Fit indices for the latent class models estimated with the Virtual Reality data ($n = 108$).

| Number of classes | LL | BIC | AIC | AIC3 | MAX. BVR | VLMR | ENTROPY R^2 |
|-------------------|----------|---------|---------|---------|----------|-----------|---------------|
| 1 | -1896.09 | 3993.52 | 3878.19 | 3921.19 | 88.10 | | |
| 2 | -1717.78 | 3697.76 | 3547.56 | 3603.56 | 72.85 | 356.63*** | 0.94 |
| 3 | -1656.94 | 3636.94 | 3451.88 | 3520.88 | 19.20 | 121.69*** | 0.93 |
| 4 | -1601.63 | 3587.19 | 3367.25 | 3449.25 | 21.36 | 110.62*** | 0.94 |
| 5 | -1564.32 | 3573.43 | 3318.63 | 3413.63 | 19.70 | 74.62*** | 0.94 |
| 6 | -1539.31 | 3584.29 | 3294.62 | 3402.62 | 14.45 | 50.01** | 0.95 |

Note: LL = log-likelihood; BIC = Bayesian information criteria; AIC = Akaike information criteria; AIC3 = AIC with 3 as penalized factor; Max. BVR = maximum of bivariate residual; VLMR = Voughn-Lo-Mendell-Rubin.
 * $p < .05$, ** $p < 0.01$, *** $p < 0.001$.

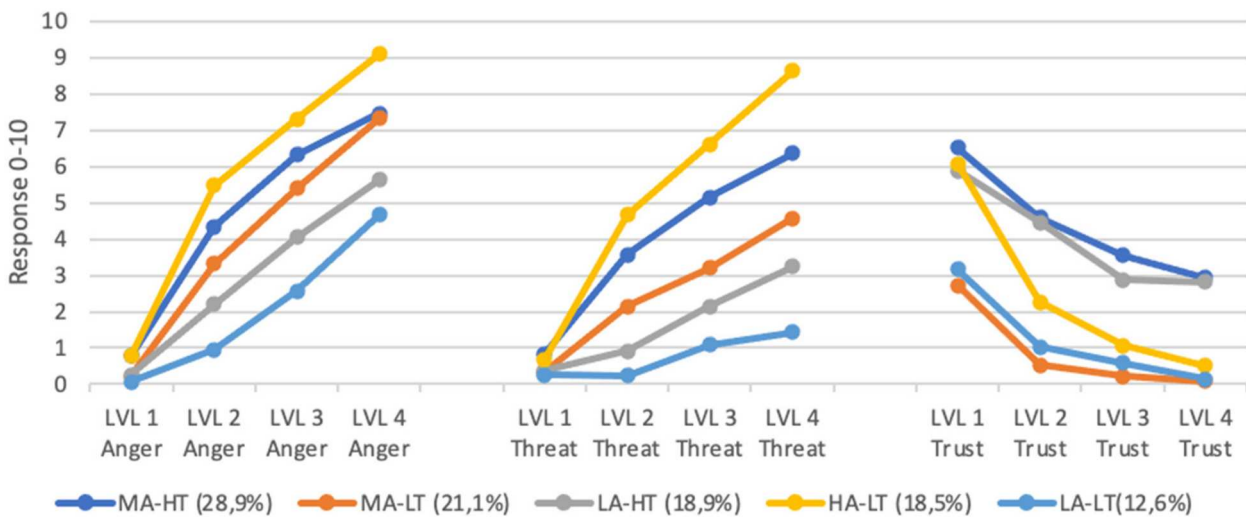


Figure 4. Profiles of state-mentalizing in Virtual Reality-task. Note: MA-HT = medium arousal, high trust; MA-LT = medium arousal, low trust; LA-HT = low arousal, high trust; HA-LT = high arousal, low trust; LA-LT = low arousal, low trust.

Table VIII. Overall and class-specific means of the 12 VR indicators, overall Wald tests, and significant pairwise Wald tests for the 5 VR-profiles ($N = 108$).

| | Overall mean | Class 1 (MA-HT) | Class 2 (MA-LT) | Class 3 (LA-HT) | Class 4 (LA-HT) | Class 5 (LA-HT) | Wald | Paired comparisons |
|-------------|--------------|-----------------|-----------------|-----------------|-----------------|-----------------|----------|---|
| LVL1 Anger | 0.47 | 0.78 | 0.21 | 0.27 | 0.77 | 0.07 | 10.80* | 2 < 1,4 |
| LVL2 Anger | 3.50 | 4.34 | 3.32 | 2.21 | 5.47 | 0.96 | 31.73*** | 2,3,5 < 1; 1,2,3,5 < 4; 5 < 2,3; 3 < 2 |
| LVL3 Anger | 5.41 | 6.32 | 5.39 | 4.05 | 7.30 | 2.57 | 30.31*** | 2,3,5 < 1; 1, 2,3,5 < 4; 3,5 < 2; 5 < 3 |
| LVL4 Anger | 7.03 | 7.46 | 7.31 | 5.62 | 9.08 | 4.67 | 36.44*** | 3,5 < 1; 1,2,3,5 < 4; 3,5 < 2; 5 < 3 |
| LVL1 Threat | 0.54 | 0.83 | 0.33 | 0.37 | 0.67 | 0.26 | 7.63 | 2 < 1 |
| LVL2 Threat | 2.55 | 3.57 | 2.14 | 0.92 | 4.67 | 0.24 | 31.26*** | 2,3,5 < 1; 1,2,3,5 < 4; 3,5 < 2; 5 < 3 |
| LVL3 Threat | 3.93 | 5.15 | 3.21 | 2.14 | 6.61 | 1.09 | 35.32*** | 2,3,5 < 1; 1,2,3,5 < 4; 3,5 < 2; 5 < 3 |
| LVL4 Threat | 5.19 | 6.36 | 4.56 | 3.25 | 8.62 | 1.43 | 36.41*** | 2,3,5 < 1; 2,3,5 < 4; 5 < 2,3 |
| LVL1 Trust | 5.09 | 6.51 | 2.73 | 5.88 | 6.07 | 3.16 | 26.68*** | 5,2 < 1; 2 < 3,4; 5 < 3,4 |
| LVL2 Trust | 2.83 | 4.59 | 0.52 | 4.46 | 2.26 | 1.03 | 26.78*** | 2,4,5 < 1; 2 < 3,4; 4,5 < 3; 5 < 4 |
| LVL3 Trust | 1.89 | 3.55 | 0.23 | 2.88 | 1.07 | 0.58 | 28.36*** | 2,4,5 < 1; 2 < 3,4; 4,5 < 3 |
| LVL4 Trust | 1.51 | 2.94 | 0.09 | 2.81 | 0.51 | 0.14 | 22.60*** | 2,4,5 < 1; 2 < 3,4; 4,5 < 3 |

Note: * $p < .05$, *** $< .001$; paired comparisons with $p < 0.05$.

anger, threat, and trust for level of aggression ranging from 1 to 4. In all classes the level of perceived anger and threat increased, and level of trustworthiness decreased when the aggressive cues increased. However, classes showed substantial differences on the level of experienced anger, threat and assessed trust. Class 1 (28.9% of the sample, $n = 31$) represents patients who experienced a medium level of anger and threat. In contrast, they perceived others as relatively trustworthy across all levels of aggressive interaction. This class is described as the Medium Aroused-High Trust profile (MA-HT-profile). The second class consisted of 21.1% of the sample ($n = 23$) and represents patients with a medium level of experienced anger and threat. However, they displayed very low trust in their counterparts in social interaction, even in the absence of, or with subtle signs of, aggression (named Medium Aroused-Low Trust-profile (MA-LT-profile)). The third class (18.9% of the sample, $n = 20$) consisted of patients with a relatively low experienced level of anger and threat. In line with this, the level of assessed trust was relatively high (named the Low Aroused – High Trust profile (LA-HT-profile)). The fourth class consisted of 18.5% of the sample ($n = 20$) with remarkable high levels of perceived anger and threat. Even when only subtle aggressive cues (like mild irritation in a person’s voice) were presented, the experienced anger and threat was still medium. Their assessed trust in social interaction was relatively high when there were no aggressive cues, however their trust in others decreased fast when aggressive cues increased (named High Aroused – Low Trust (HA-LT-profile)). The last class (12.6%, $n = 14$) exhibited

a pattern of very low perceived anger and threat in social interaction, even in the presence of high levels of aggressive behavior. This group displayed, however, little trust in their counterparts (named Low Aroused – Low Trust-profile (LA-LT-profile)).

Exploration of Relations Between Instruments of Mentalizing Capacities

A step-3 analysis was conducted to examine whether trait mentalizing (RFS and ERART) was related to state mentalizing (VR). First, the relationship between the class membership of the ERART and the VR was investigated. There was no significant relationship between class memberships of both instruments (Wald = 4.63, $p = .8$). The conditional probability of the VR-class membership given ERART-class membership can be found in Table IX. Secondly, the relationship between the overall capacity to mentalize, measured with the RFS based on the Adult

Table IX. Conditional probabilities of belonging to an ERART-class conditional on VR-class membership.

| | Class 1 (MA-HT) | Class 2 (MA-LT) | Class 3 (LA-HT) | Class 4 (HA-LT) | Class 5 (LA-LT) |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| High-ER-profile | 0.55 | 0.58 | 0.58 | 0.39 | 0.38 |
| INE-profile | 0.38 | 0.37 | 0.31 | 0.60 | 0.54 |
| Low-ER-profile | 0.07 | 0.06 | 0.11 | 0.01 | 0.08 |

Table X. Mean score of RFS and false positives on the ERART indicator ‘anger’ per VR class.

| | <i>Class 1 (MA-HT)</i> | <i>Class 2 (MA-LT)</i> | <i>Class 3 (LA-HT)</i> | <i>Class 4 (LA-HT)</i> | <i>Class 5 (LA-LT)</i> | <i>Wald</i> | <i>p-value</i> | <i>Paired comparisons</i> |
|----------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|-------------|----------------|---------------------------|
| Mean RFS-score | 3.14 | 2.50 | 3.15 | 2.76 | 3.26 | 1.54 | .82 | |
| Anger bias | 1.99 | 2.86 | 1.92 | 3.77 | 2.37 | 12.06 | .02 | 1 < 4; 3 < 4 |

Table XI. Mean score of RFS and false positives on the ERART indicator ‘anger’ per ERART class.

| | <i>Class 1 (High-ER-profile)</i> | <i>Class 2 (INE-profile)</i> | <i>Class 3 (Low-ER-profile)</i> | <i>Wald</i> | <i>p-value</i> | <i>Paired comparisons</i> |
|----------------|--------------------------------------|----------------------------------|-------------------------------------|-------------|----------------|---------------------------|
| Mean RFS-score | 3.19 | 2.99 | 2.60 | 0.41 | .81 | |
| Anger bias | 1.36 | 4.02 | 1.86 | 27.03 | <.001 | 1 < 2; 3 < 2 |

Attachment Interview, and the class memberships of the ERART and VR were estimated. The overall relationship between the RFS-score and the VR-class membership was not significant (Wald = 1.54, $p = .82$). Also, the overall relationship between the RFS-score and ERART-class membership was not significant (Wald = 0.41, $p = .81$). Class-specific means of the RFS for ERART and VR classes are presented in [Tables X](#) and [XI](#), respectively.

Thirdly, the associations between on the one hand ERART and VR class memberships and on other hand anger bias were investigated. Note that the false positive scores of anger from the ERART were used to measure hostile attribution bias. The relationship between anger bias and VR class membership was significant (Wald = 12.06, $p = .02$). The HA-LT class had significant higher scores on anger bias than the MA-HT and LA-HT classes. Regarding the ERART, a significant relationship was found between anger bias and the ERART class memberships (Wald = 27.03, $p = <.001$; Wald = 23.14, $p = <.001$). Patients in the INE-profile had significant higher anger bias-scores than patients in the High-ER and Low-ER profiles. The class-specific means of the bias scores for the ERART and VR classes are presented in [Tables X](#) and [XI](#), respectively.

Discussion

In this study, we investigated whether there are different mentalizing profiles within a population of patients with antisocial behavior. Various instruments for measuring mentalizing capacity were used to assess both trait- and state-mentalizing. Latent class analyses were employed to determine whether distinct mentalizing profiles could be identified. Additionally, the relationship between trait- and state mentalizing was examined. As expected, and in line with previous studies (Abi-Habib et al.,

2020; Newbury-Helps et al., 2017; Protic et al., 2020), we found reduced mentalizing capacities in patients with antisocial behavior. These reduced capacities were apparent in various aspects of the mentalizing spectrum, like reflective functioning, the recognition of emotions and state-mentalizing.

With regard to reflective functioning, a mean RFS-score of 2.97 was found with a range of -1 to 6 (maximum range is -1 to 9). This is consistent with findings from other studies in comparable samples. For instance, a mean RFS-score of 2.65 was found in a clinical group of youth (Taubner et al., 2016) and a score of 2.11 in a sample of prisoners (Fonagy & Levinson, 2004). The average score in the current study is slightly higher, which can be explained by the broad research population, ranging from individuals with low levels of antisocial behavior to those with high levels of antisocial behavior. Noteworthy is that, despite the relatively low score for mentalizing capacity, there is still some degree of variation. Mentalizing capacities range from negative/low to average. This may be indicative for the heterogeneity in mentalizing capacity among individuals with antisocial behavior conform our hypothesis.

With regard to emotion recognition, the level of recognition of fear and sadness is very low. Although difficulties with recognizing fear and sadness align with previous research findings in healthy populations (Kessels et al., 2014; Montagne et al., 2005; Montagne et al., 2007), the level of emotion recognition of fear and sadness compared to the health populations is very low for all profiles in the current sample, especially for the INE- and low ER-profiles. These results are in line with research outcomes of the meta-analysis of Marsh and Blair (2008), who found that antisocial individuals were particularly limited in recognizing fear and sadness compared to control groups.

Besides general low recognition of fear and sadness, there is heterogeneity in the recognition of other emotions. Our results also show that there is no uniform profile of emotion recognition in the current sample, which indicates, in line with our hypothesis, that there is heterogeneity in emotion recognition capacity. Three different emotion recognition-profiles were identified with a high degree of class membership predictability. A large part of the participants (50.5%, high-ER-profile) is able to adequately recognize the emotions anger, surprise, disgust and happiness. The recognition accuracy is comparable to healthy populations on the same stimuli (Kessels et al., 2014; Montagne et al., 2005; Montagne et al., 2007). The INE-profile (42.1%), however, is characterized by specific problems with recognizing the negative emotions fear, sadness and disgust. Although participants within this profile recognize the emotions anger and surprise more accurately, they perform less accurate than the high-ER-profile. Interestingly, Kleine Deters et al. (2022) found similar profiles in their study within a combined sample (healthy and clinical) of adolescents. They found relatively low recognition of fear, sadness, and disgust in the cluster characterized by reactive aggression and high CU-traits. The results within the INE-profile are also partly in line with the systematic review of Chapman et al. (2018), where violent offenders show impairments in recognizing fear and disgust. The third profile, the low-ER-profile, is characterized by low recognition of all emotions. However, this profile consists of a small group of the sample (7.4%). The level of accurate emotion recognition of individuals within this profile is lower than in comparable studies (Kosson et al., 2006; Montagne et al., 2005). Inaccurate recognition of both positive emotions (surprise and happiness) and negative emotions in this cluster has also been found by other researchers. For example, Dawel et al. (2012) found support for difficulties in emotion recognition for both positive and negative emotions, including fear, happiness and surprise in their meta-analysis of psychopathic samples. Similarly, Marsh and Blair (2008) found emotion recognition problems for fear, sadness and surprise.

Given that hostile attribution bias is a well-documented issue in mentalizing among individuals with antisocial problems, further investigation was conducted into this bias within the context of emotion recognition. Notably, individuals with specific difficulties in recognizing negative emotions (INE-profile) scored significantly higher on anger bias compared to the other two groups (high-ER and low-ER). Misinterpretation of other (ambiguous) emotions as if it were anger, has also been demonstrated in previous research (Mellentin et al.,

2015; Schonenberg & Jusyte, 2014; Smeijers et al., 2017). The INE-profile shows that there is not a general difficulty in emotion recognition, but rather a misinterpretation and heightened focus on hostility or anger in others, which may also contribute to the misidentification of other (negative) emotions as anger. The various deficits in emotion recognition within the profiles underscore the importance of mapping specific mentalizing shortcomings in this population. An important finding is that in general, the distress-stimuli fear and sadness are poorly recognized. However, there appears to be variation within our sample in the extent to which other emotions, like anger, happiness, surprise, and disgust, are recognized accurately. Also, a specific problem with a hostile attribution bias is evident for some individuals.

In addition to the ability to recognize emotions, being able to adequately interpret the mental states of oneself and others in social interaction is also an important aspect of mentalizing. While mentalizing is mostly operationalized in research as a stable capacity (trait-mentalizing), theories about mentalizing suggest that mentalizing capacity may change under contextual factors, for example high arousal (Bateman & Fonagy, 2012). This mentalizing capacity is also known as state-mentalizing. Problems in state-mentalizing, for example a disbalance between controlled and automatic mentalizing in direct social interaction, are mentioned by several researchers as possible mentalizing problem in anti-social populations (Bateman & Fonagy, 2012; Fonagy & Luyten, 2018). To gain insight into state-mentalizing, a Virtual Reality experiment was conducted in the current study. In the virtual environment, situations were created with increasing levels of hostile ostensive cues in social interactions. In these scenarios, we examined participants' perceptions of anger and threat in their interaction with others, as well as the extent in which they found the other person trustworthy.

Regarding the virtual reality experiment, it was assumed that with adequate mentalizing capacities, increasing aggression leads to higher levels of perceived anger and threat in a proportional way as well as decreased levels of trustworthiness. The avatar can display four levels of aggression, ranging from 0 (no aggression, neutral conversation) to 4 (high level of aggression). The first observation that emerged clearly was that there was significant variation in the experience of anger and perceived threat in interaction with others. Notably, the LA-HT and LA-LT groups reported lower levels of anger and perceived threat. Conversely, the HA-LT group interpreted slight signals of hostility from the other party as angry and threatening. Groups

appear to differ, firstly, in the extent to which they incorporate ostensive cues in their assessment of others. The LA-HT and LA-LT groups seem to integrate these cues only to a limited extent, whereas the HA-LT group does so to a considerable degree. The variation among the profiles may also be explained by diminished affective mentalizing in patients who experience low arousal in interactions with others. The emotional impact of the other person's behavior or intentions is felt to a lesser extent in this context. In line with this, it is also known that many people with antisocial behavioral problems live in hostile environments (Burnside & Gaylord-Harden, 2019) and may therefore be used to aggressive interactions and not easily be intimidated by them. On the other hand, an imbalance between automatic and controlled mentalizing may also play a role in the experience of high arousal (as seen in the HA-LT group), where automatic thoughts about the other party are not corrected by controlled mentalizing.

The second important observation is that (interpersonal) mistrust appears to play a significant role in social interaction for a large portion of the study sample. More than 50% of the participants (LA-LT, MA-LT, and HA-LT-group) experienced others as not or only somewhat trustworthy when confronted with even subtle signs of hostility. Moreover, 33.7% (LA-LT and MA-LT-group) experienced foundational low (interpersonal) trust even in friendly conversations, suggesting ostensive cues of trustworthiness are missed or not integrated when appraising a situation. This supports the work of Bateman and Fonagy (2012), who identify epistemic trust as the foundation for developing adequate mentalizing capacities. The absence or underdevelopment of interpersonal trust can lead to chronic distrust in others. These findings align with the imbalance between automatic and controlled mentalizing as proposed by Bateman and Fonagy (2012) and Gagliardini et al. (2023), where automatic assumptions are not corrected by controlled mentalizing. These participants seem to be unable to trust other persons and are reluctant to modify their ideas about others, even when the context is friendly and supportive. They show rigid patterns of mentalizing that can be seen by traumatized individuals (Benzi et al., 2023).

The other 47.8% of the participants (MA-HT and LA-HT-group) show relative high levels on perceived trustworthiness. Although trustworthiness decreases with increasing hostile interaction, it remains relatively high even during very aggressive interactions. A possible explanation for this high levels of trustworthiness is an imbalance in internal and affective mentalizing, where there is a limited reflection on one's own experiences in social

interaction and insufficient engagement with the emotional values of the interaction. Another explanation is, that these people have a more rigid pattern of mentalizing. This may be in line with earlier findings about rigid mentalizing modes by Benzi et al. (2023), where individuals accept the other person without critical reflection. This is known as epistemic credulity and associated with traumatic experiences of emotional abuse. Epistemic credulity makes individual susceptible for exploitation and misinformation (Benzi et al., 2023).

The VR-profiles were also compared with the level of hostile attribution bias in the emotion recognition-task (ERART). An important finding was that individuals in the HA-LT profile exhibited significantly higher hostile attribution bias than those in the MA-HT and LA-HT profiles. Schonenberg and Jusyte (2014) previously found that individuals with antisocial behavior not only tend to interpret emotions as hostile but also overestimate the intensity of perceived anger. This pattern is also reflected in the HA-LT profile, where both anger and perceived threat are rated very high, even in response to subtle signs of hostility. Additionally, this profile shows that perceived trustworthiness declines rapidly in the presence of these subtle hostile cues. The evaluation of the counterpart's trustworthiness for individuals in this profile may be affected by a hostile attribution bias. Further research into the role of hostile attribution bias in relation to interpersonal trust is warranted.

Considering the findings from the VR, we observe an important role for interpersonal trust in state-mentalizing combined with imbalances in automatic and affective mentalizing, which aligns with previously proposed models by Bateman et al. (2013); Fonagy and Luyten (2018), and Gagliardini et al. (2023). Also, a hostile attribution bias seems to have a role in (mis)interpreting external cues in direct social interaction for a part of the sample. Findings indicate that individuals with antisocial behavior display deficiencies in mentalizing capacities across various dimensions. It is therefore essential to further map these different deficiencies through experimental research.

Another aim of the current study was to investigate the relationship between trait-mentalizing and state-mentalizing. The RFS and ERART measure dimensions of mentalizing as stable trait, not including direct social interaction. The VR-experiment, however, includes direct social interaction and is therefore more indicative as measure for state-mentalizing. We found no significant relation between the profiles on the ERART and VR-task, suggesting that emotion recognition profiles are not indicative for specific profiles of mentalizing in social

interaction (VR-task) and vice versa. Furthermore, there was also no significant relation between general reflective functioning (RFS-score) and the profiles of the ERART and VR-task. This confirms our hypothesis that trait- and state-mentalizing are not necessarily related. A possible explanation for the lack of relation between these instruments is that these instruments measure distinct aspects of mentalizing. Another explanation is that there are within-person fluctuations of mentalizing (Steinberg et al., 2023). In empirical research, mentalizing is often measured as a fixed capacity. However, individuals may fluctuate in their capacity to mentalize when contexts change. For example, some individuals may have more problems with balancing mentalization under high arousal (Luyten et al., 2020; Nolte et al., 2013). The RFS measures mentalization as a stable personality trait (Hörz-Sagstetter et al., 2015), which cannot adequately capture fluctuations. The VR-task, however, relies more on state-related mentalizing, where contextual factors (such as increasing arousal) can influence mentalizing capacities at a specific moment. This allows for the assessment of moment-specific mentalizing in individuals, particularly of those who are more sensitive to emotion dysregulation. Functioning on the RFS may in that case not be related to state-mentalizing on the VR-task. However, although the profiles of the ERART and VR do not generally correlate, we did find that a hostile attribution bias, measured with the ERART, did differentiate between the VR profiles. From this, we can conclude that specific issues in emotion recognition may influence state-mentalizing modes.

Clinical Implications

Research findings show heterogeneity in mentalizing abilities in individuals with antisocial behavior. This advocates for thorough diagnostics in individuals with antisocial problems, mapping out specific dysfunctions in mentalizing on an individual level. It is important to distinguish between mentalization as personality trait, which is relatively stable over time, and mentalizing in direct interaction with others. In diagnostics, it is crucial to use assessment tools that capture both aspects, to obtain a clear understanding of an individual's underlying difficulties in mentalizing. For instance, a person might exhibit few problems with mentalizing in situations with low stress but temporarily lose the ability to mentalize under high arousal. Contrary, other individuals may experience more stable mentalizing problems, even under low arousal. Accurate diagnostics in this regard can provide targeted indications for treatment interventions.

If careful assessment of mentalizing difficulties is conducted, treatment can also be more specifically tailored to address these mentalizing deficits. The latent classes indicate different mentalizing difficulties, each of which requires a distinct treatment focus. Generally, low mentalizing capacities are observed in this study among individuals with antisocial behavior. In line with the recommendation for therapeutic interventions for patients with ASPD, Mentalizing Based Treatment (MBT) seems meaningful to apply to individuals with antisocial behavioral problems as well, specifically for individuals with low mentalizing capacities (hypomentalizing). Within the population of individuals with comorbid borderline personality disorder and ASPD, MBT has proven to have effect on the reduction of, for example, anger and hostility, and self-harming behavior (Bateman et al., 2016). Also for individuals with conduct disorder, MBT supports effective mentalizing (Kasper et al., 2024). It is recommended to further investigate the efficacy of MBT within a population of individuals with antisocial behavior. Furthermore, it is notable that within the latent classes, some individuals tend to rely more on automatic mentalizing, drawing conclusions about the other without verifying them. Examples include overreporting perceived anger and threat, distrusting the other party without any clear justification, and anger biases. For this group, interventions such as Schema-Focused Therapy or Cognitive Behavioral Therapy, which challenge negative cognitions, may be more appropriate.

Furthermore, more than 50% of the participants with antisocial behavior showed problems with their experienced trustworthiness of others. This indicates an important role for interpersonal trust in the psychotherapeutic process. Psychotherapy should focus on teaching patients with antisocial behavior to include different sources of information when determining the trustworthiness of other people. The issue of interpersonal trust also plays a role in the therapeutic alliance and managing trust issues in the therapeutic alliance may be a therapeutic mean in itself (Aerts et al., 2023). This will take time, effort and patience, as with interpersonal trust issues, increasing reflective functioning also needs work. The results of several studies, however, suggest that insight in specific mentalizing characteristics may increase therapists' capacity and motivation to provide psychotherapy to these patients (Flaaten et al., 2024; Van Dam et al., 2022).

Strengths and Limitations

This study has several strengths. Previous research on mentalizing functioning has been conducted in diverse

populations (offenders, individuals with ASPD, healthy populations). In the current study, a broader perspective on antisocial behavior is adopted, including a transdiagnostic sample from both general- and forensic mental health settings. This aligns with trends towards transdiagnostic thinking and addresses the prevalence of antisocial issues within broader mental health care. Additionally, this study utilizes a combination of research instruments, including experimental tasks. This enhances the reliability of the results compared to studies relying solely on self-report, which may be less suitable given the nature of this population's antisocial issues. The use of different instruments also makes it possible to gain more insight into both trait-mentalizing and state-mentalizing. Previous research was often focused on mentalizing as fixed capacity and less on within-person fluctuations in mentalizing. The approach of the current study enhances insight in mentalizing impairments on both trait- and state-level and underlines the importance of thorough examination of mentalizing problems in individuals.

Although the use of a broad research population of individuals with antisocial problems has advantages, it also presents limitations. The broad sample may include greater variation in mentalizing problems compared to an ASPD-specific sample. This limits the generalizability to a specific ASPD target group. However, given the similarities in mentalizing abilities, the proposed models for ASPD and the accompanying psychotherapies may therefore also be applicable for antisocial behavior in general. Additionally, the relatively small sample size is a limitation of this study. Nylund-Gibson and Choi (2018) mentioned a range of 300–1000 as sample-size indication for LCA, however, they also stated that small sample sizes, for example 30, may be sufficient for models with well-separated classes. The sample size ($n = 108$) of the current study could potentially reduce the power to identify smaller classes or to distinguish between classes. However, the current findings are consistent with previous research on mentalizing capacities. Furthermore, specifically regarding the sample size of the RFS, there was no indication of a power issue. However, further research in larger samples is recommended. Furthermore, the current study employs stimuli in the emotion recognition task varying in clarity levels (30% to 100% clarity). To minimize participants' burden, each stimulus is presented only once. This approach may have affected the reliability of the subscales. However, the primary focus of this study is the reliability of the latent classes. This is reflected in the entropy R^2 , which demonstrates a strong predictive value. The instruments are capable of accurately predicting class assignment, indicating its effectiveness in distinguishing between classes.

The latent classes align with theoretical expectations, reinforcing the validity of the instruments. Regarding the RFS it is worth noticing that the coders were not blinded to the research questions, which may have influenced the coding process. The coders were blind to the latent classifications and results on other instruments, though. Besides, there was a high inter-rater reliability between the coders. Other limitations are associated with the use of the virtual reality program. In addition to the benefits of using real-life interaction through virtual reality, there are also drawbacks associated with this method. Technical constraints of the program used, such as standardized speech text and the presentation of a limited number of virtual characters, may have influenced the response of the participants. It is also the first time that this program is used as assessment tool in scientific research. It remains unclear to what extent the results can be generalized to real-life social interactions. Further research is needed to explore this issue in more detail. Additionally, greater attention should be given to the operationalization of the concept of "trust". In literature various terms, such as trust, epistemic trust, interpersonal trust, and perceived/expected trustworthiness are used. Further research is needed to effectively measure the concept of trust in direct social interactions, as well as epistemic credibility. This could involve combining questionnaires on trust with experiments (such as virtual reality), where trust and related concepts are measured in direct contact with others. Despite the limitations of using this method of measurement, the experiment demonstrates that it is capable of making an accurate prediction of class assignment which indicates that the tool is capable of effectively differentiating between classes. Further research is, however, necessary to evaluate the use of virtual reality as a measurement instrument. Finally, this study did not include a comparison with normative data on, for example, emotion recognition. This is recommended for future research.

Conclusion

Our results show that the majority of patients with antisocial behavior who participated in this study displayed reduced mentalizing capacities. Besides poor general mentalizing capacities, there is also evidence for difficulties with emotion recognition, specifically for the emotions of fear and sadness. However, individuals differ on the level in which they recognize other emotions, such as anger, surprise, disgust, and happiness. Mentalizing deficits were also observed in the domain of state-mentalizing. In direct social interaction, 52.5% of participants reported difficulties with trusting others, suggesting an imbalance between automatic and controlled

mentalizing. Additionally, some patients exhibited reduced emotional reactivity to stimuli, indicating deficits in affective mentalizing. Furthermore, there were indications of a hostile attribution bias in a subset of participants. This bias was associated with increased feelings of anger and threat, and reduced trust in direct social interactions. Our findings provide support for the conceptual models proposed for ASPD and conduct problems (Bateman et al., 2013; Fonagy & Luyten, 2018). However, this study also highlights the heterogeneity within the population of individuals with antisocial behavior and the importance of mapping mentalizing deficits on an individual level.

Author Contributions

BW, AD, and MR designed the study. BW and MH collected the data. BW and JV did the analyses. BW drafted the first version of the manuscript. MR, AD, JV, and MH reviewed and provided critical revisions to the manuscript. All authors approved the final version for publication.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

ORCID

Brenda De Wit-De Visser  <http://orcid.org/0000-0001-7621-7001>

Madeleine J.N. Rijckmans  <http://orcid.org/0000-0001-6486-632X>

Jeroen K. Vermunt  <http://orcid.org/0000-0001-9053-9330>

Matthijs J.W. Hamakers  <http://orcid.org/0009-0002-7675-1356>

Arno Van Dam  <http://orcid.org/0000-0002-5604-9837>

Data Availability

The data used for this research is available on request from the corresponding author. Due to ethical restrictions, the data are not publicly available.

References

Abi-Habib, R., Wehbe, N., Badr, K., & Tohme, P. (2020). Do prisoners mentalize differently? Investigating attachment and reflective functioning in a sample of incarcerated Lebanese

- men. *International Journal of Forensic Mental Health*, 19(2), 183–197. <https://doi.org/10.1080/14999013.2019.1684403>
- Aerts, J. E. M., Rijckmans, M. J. N., Bogaerts, S., & van Dam, A. (2023). Establishing an optimal working relationship with patients with an antisocial personality disorder: Aspects and processes in the therapeutic alliance. *Psychology and Psychotherapy: Theory Research and Practice*, 96(4), 999–1014. <https://doi.org/10.1111/papt.12492>
- Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (1993). A psychometric study of the adult attachment interview: Reliability and discriminant validity. *Developmental Psychology*, 29(5), 870–879. <https://doi.org/10.1037/0012-1649.29.5.870>
- Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling*, 23(1), 20–31. <https://doi.org/10.1080/10705511.2014.955104>
- Bateman, A., Bolton, R., & Fonagy, P. (2013). Antisocial personality disorder: A mentalizing framework. *FOCUS*, 11(2), 178–186. <https://doi.org/10.1176/appi.focus.11.2.178>
- Bateman, A., & Fonagy, P. (2012). *Handbook of mentalizing in mental health practice*. American Psychiatric Pub.
- Bateman, A., & Fonagy, P. (2016). *Mentalization-based treatment for personality disorders: A practical guide*. Oxford University Press. <https://doi.org/10.1093/med/psych/9780199680375.001.0001>
- Bateman, A., Fonagy, P., & American Psychiatric Association, P. (2019). *Handbook of mentalizing in mental health practice* (Second edition, ed.). American Psychiatric Association Publishing. <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2171163>. <https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5790646>.
- 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. <https://doi.org/10.1186/s12888-016-1000-9>
- Benzi, I. M. A., Carone, N., Parolin, L., Martin-Gagnon, G., Ensink, K., & Fontana, A. (2023). Different epistemic stances for different traumatic experiences: Implications for mentalization. *Research in Psychotherapy*, 26(3), 708. <https://doi.org/10.4081/ripppo.2023.708>
- Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition*, 57(1), 1–29. [https://doi.org/10.1016/0010-0277\(95\)00676-P](https://doi.org/10.1016/0010-0277(95)00676-P)
- Blair, R., Colledge, E., Murray, L., & Mitchell, D. (2001). A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Child Psychology*, 29(6), 491–498. <https://doi.org/10.1023/A:101225108281>
- Bolck, A., Croon, M., & Hagenars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1), 3–27. <https://doi.org/10.1093/pan/mph001>
- Burnside, A. N., & Gaylord-Harden, N. K. (2019). Hopelessness and delinquent behavior as predictors of community violence exposure in ethnic minority male adolescent offenders. *Journal of Abnormal Child Psychology*, 47(5), 801–810. <https://doi.org/10.1007/s10802-018-0484-9>
- Burt, S. A., & Donnellan, M. (2009). Development and validation of the subtypes of antisocial behavior questionnaire. *Aggressive Behavior*, 35(5), 376–398. <https://doi.org/10.1002/ab.20314>
- Chapman, H., Gillespie, S. M., & Mitchell, I. J. (2018). Facial affect processing in incarcerated violent males: A systematic review. *Aggression and Violent Behavior*, 38, 123–138. <https://doi.org/10.1016/j.avb.2017.10.006>

- Cima, M., Raine, A., Meesters, C., & Popma, A. (2013). Validation of the Dutch reactive proactive questionnaire (RPQ): differential correlates of reactive and proactive aggression from childhood to adulthood. *Aggressive Behavior, 39*(2), 99–113. <https://doi.org/10.1002/ab.21458>
- Dawel, A., O’Kearney, R., McKone, E., & Palermo, R. (2012). Not just fear and sadness: Meta-analytic evidence of pervasive emotion recognition deficits for facial and vocal expressions in psychopathy. *Neuroscience and Biobehavioral Reviews, 36*(10), 2288–2304. <https://doi.org/10.1016/j.neubiorev.2012.08.006>
- De Castro, B. O., Veerman, J. W., Koops, W., Bosch, J. D., & Monshouwer, H. J. (2002). Hostile attribution of intent and aggressive behavior: A meta-analysis. *Child Development, 73*(3), 916–934. <https://doi.org/10.1111/1467-8624.00447>
- De Wit-De Visser, B., Rijckmans, M., Vermunt, J. K., & van Dam, A. (2023). Pathways to antisocial behavior: A framework to improve diagnostics and tailor therapeutic interventions [hypothesis and theory]. *Frontiers in Psychology, 14*, 993090. <https://doi.org/10.3389/fpsyg.2023.993090>
- Dodge, K. A., & Crick, N. R. (1990). Social information-processing bases of aggressive behavior in children. *Personality and Social Psychology Bulletin, 16*(1), 8–22. <https://doi.org/10.1177/0146167290161002>
- Flaaten, E., Langfeldt, M., & Morken, K. T. E. (2024). Antisocial personality disorder and therapeutic pessimism – how can mentalization-based treatment contribute to an increased therapeutic optimism among health professionals? *Frontiers in Psychology, 15*, 1320405. <https://doi.org/10.3389/fpsyg.2024.1320405>
- Fonagy, P., & Levinson, A. (2004). Offending and attachment: The relationship between interpersonal awareness and offending in a prison population with psychiatric disorder. *Canadian Journal of Psychoanalysis, 12*(2), 225–251. <https://doi.org/10.1002/j.2167-4086.2009.tb00406.x>
- Fonagy, P., & Luyten, P. (2018). Conduct problems in youth and the RDoC approach: A developmental, evolutionary-based view. *Clinical Psychology Review, 64*, 57–76. <https://doi.org/10.1016/j.cpr.2017.08.010>
- Fonagy, P., Luyten, P., Moulton-Perkins, A., Lee, Y.-W., Warren, F., Howard, S., Ghinai, R., Fearon, P., & Lowyck, B. (2016). Development and validation of a self-report measure of mentalizing: The reflective functioning questionnaire. *PLoS One, 11*(7), e0158678. <https://doi.org/10.1371/journal.pone.0158678>
- Gagliardini, G., Gullo, S., Teti, A., & Colli, A. (2023). Personality and mentalization: A latent profile analysis of mentalizing problematics in adult patients. *Journal of Clinical Psychology, 79*(2), 514–530. <https://doi.org/10.1002/jclp.23430>
- Gergely, G., & Unoka, Z. (2008). Attachment and mentalization in humans: The development of the affective self. In E. L. Jurist, A. Slade, & S. Bergner (Eds.), *Mind to mind: Infant research, neuroscience, and psychoanalysis* (pp. 50–87). Other Press.
- Hörz-Sagstetter, S., Mertens, W., Isphording, S., Buchheim, A., & Taubner, S. (2015). Changes in reflective functioning during psychoanalytic psychotherapies. *Journal of the American Psychoanalytic Association, 63*(3), 481–509. <https://doi.org/10.1177/0003065115591977>
- Kasper, L., Hauschild, S., Schrauf, L., & Taubner, S. (2024). Enhancing mentalization by specific interventions within mentalization-based treatment of adolescents with conduct disorder. *Frontiers in Psychology, 14*, 1223040. <https://doi.org/10.3389/fpsyg.2023.1223040>
- Kessels, R. P. C., Montagne, B., Hendriks, A. W., Perrett, D. I., & de Haan, E. H. F. (2014). Assessment of perception of morphed facial expressions using the emotion recognition task: Normative data from healthy participants aged 8–75. *Journal of Neuropsychology, 8*(1), 75–93. <https://doi.org/10.1111/jnp.12009>
- Kleine Deters, R., Naaijen, J., Holz, N. E., Banaschewski, T., Schulze, U. M. E., Sethi, A., Craig, M. C., Sagar-Ouriaghli, I., Santosh, P., Rosa, M., Castro-Fornieles, J., Penzol, M. J., Arango, C., Brandeis, D., Franke, B., Glennon, J. C., Buitelaar, J. K., Hoekstra, P. J., & Dietrich, A. (2022). Emotion recognition profiles in clusters of youth based on levels of callous-unemotional traits and reactive and proactive aggression. *European Child & Adolescent Psychiatry, 32*(12), 2415–2425. <https://doi.org/10.1007/s00787-022-02079-3>
- Klein Tuentje, S., Bogaerts, S., & Veling, W. (2019). Hostile attribution bias and aggression in adults – a systematic review. *Aggression and Violent Behavior, 46*, 66–81. <https://doi.org/10.1016/j.avb.2019.01.009>
- Kosson, D. S., Lorenz, A. R., & Newman, J. P. (2006). Effects of comorbid psychopathy on criminal offending and emotion processing in male offenders with antisocial personality disorder. *Journal of Abnormal Psychology, 115*(4), 798–806. <https://doi.org/10.1037/0021-843X.115.4.798>
- Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., Eaton, N. R., Forbes, M. K., Forbush, K. T., Goldberg, D., Hasin, D., Hyman, S. E., Ivanova, M. Y., Lynam, D. R., Markon, K., ... Zimmerman, M. (2017). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology, 126*(4), 454–477. <https://doi.org/10.1037/abn0000258>
- 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., Campbell, C., Moser, M., & Fonagy, P. (2024). The role of mentalizing in psychological interventions in adults: Systematic review and recommendations for future research. *Clinical Psychology Review, 108*, 102380. <https://doi.org/10.1016/j.cpr.2024.102380>
- Marsh, A., & Blair, R. (2008). Deficits in facial affect recognition among antisocial populations: A meta-analysis. *Neuroscience and Biobehavioral Reviews, 32*(3), 454–465. <https://doi.org/10.1016/j.neubiorev.2007.08.003>
- Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods, 44*(2), 314–324. <https://doi.org/10.3758/s13428-011-0168-7>
- Mellentin, A. I., Stenager, E., Kirk, U., Dervisevic, A., & Pilegaard, M. (2015). Seeing enemies? A systematic review of anger bias in the perception of facial expressions among anger-prone and aggressive populations. *Aggression and Violent Behavior, 25*, 373–383. <https://doi.org/10.1016/j.avb.2015.09.001>
- Milesi, A., De Carli, P., Locati, F., Benzi, I. M. A., Campbell, C., Fonagy, P., & Parolin, L. (2023). How can I trust you? The role of facial trustworthiness in the development of epistemic and interpersonal trust. *Human Development, 67*(2), 57–68. <https://doi.org/10.1159/000530248>
- Miller, S. D., & Duncan, B. L. (2000). *The outcome rating scale*. Chicago: Author.
- Montagne, B., Kessels, R. P. C., De Haan, E. H. F., & Perrett, D. I. (2007). The emotion recognition task: A paradigm to measure the perception of facial emotional expressions at different intensities. *Perceptual and Motor Skills, 104*(2), 589–598. <https://doi.org/10.2466/pms.104.2.589-598>

- Montagne, B., van Honk, J., Kessels, R. P. C., Frigerio, E., Burt, M., van Zandvoort, M. J. E., Perrett, D. I., & de Haan, E. H. F. (2005). Reduced efficiency in recognising fear in subjects scoring high on psychopathic personality characteristics. *Personality and Individual Differences*, 38(1), 5–11. <https://doi.org/10.1016/j.paid.2004.02.008>
- Newbury-Helps, J., Feigenbaum, J., & Fonagy, P. (2017). Offenders with antisocial personality disorder display more impairments in mentalizing. *Journal of Personality Disorders*, 31(2), 232–255. https://doi.org/10.1521/pedi_2016_30_246
- Nolte, T., Bolling, D. Z., Hudac, C. M., Fonagy, P., Mayes, L., & Pelphey, K. A. (2013). Brain mechanisms underlying the impact of attachment-related stress on social cognition. *Frontiers in Human Neuroscience*, 7, 816. PMID 24348364. <https://doi.org/10.3389/fnhum.2013.00816>
- Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, 4(4), 440–461. <https://doi.org/10.1037/tps0000176>
- Protic, S., Wittmann, L., Taubner, S., & Dimitrijevic, A. (2020). Differences in attachment dimensions and reflective functioning between traumatized juvenile offenders and maltreated non-delinquent adolescents from care services. *Child Abuse & Neglect*, 103, <https://doi.org/10.1016/j.chiabu.2020.104420>
- Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., Stouthamer-Loeber, M., & Liu, J. (2006). The reactive-proactive aggression questionnaire: Differential correlates of reactive and proactive aggression in adolescent boys. *Aggressive Behavior*, 32(2), 159–171. <https://doi.org/10.1002/ab.20115>
- Rijckmans, M. J. N., Van Dam, A., & Van den Bosch, L. M. C. (2020). *Praktijkboek antisociaal gedrag en persoonlijkheidsproblematiek*. Bohn Stafleu van Loghum.
- Schonenberg, M., & Jusyte, A. (2014). Investigation of the hostile attribution bias toward ambiguous facial cues in antisocial violent offenders. *European Archives of Psychiatry and Clinical Neuroscience*, 264(1), 61–69. <https://doi.org/10.1007/s00406-013-0440-1>
- Smeijers, D., Rinck, M., Bulten, E., van den Heuvel, T., & Verkes, R. J. (2017). Generalized hostile interpretation bias regarding facial expressions: Characteristic of pathological aggressive behavior. *Aggressive Behavior*, 43(4), 386–397. <https://doi.org/10.1002/ab.21697>
- Soe-Agnie, S., Patrick, C., Nijman, H., & Jong, C. (2015). Validation of the full and brief externalizing spectrum inventory in Dutch forensic inpatients. *The Journal of Forensic Psychiatry & Psychology*, 27, 1–15. <https://doi.org/10.1080/14789949.2015.1090621>
- Steinberg, N., Moshe-Cohen, R., Sheena, L., & Kivity, Y. (2023). Capturing a mentalized moment: A pilot study of the psychometric properties of a novel assessment method of mentalizing in daily life. *Current Psychology*, 43, 1–16. <https://doi.org/10.1007/s12144-023-04963-w>
- Taubner, S., Zimmermann, L., Ramberg, A., & Schröder, P. (2016). Mentalization mediates the relationship between early maltreatment and potential for violence in adolescence. *Psychopathology*, 49(4), 236–246. <https://doi.org/10.1159/000448053>
- Van Dam, A., Rijckmans, M., & Bosch, L. (2022). Explaining the willingness of clinicians to work with patients with antisocial personality disorder using the theory of planned behaviour and emotional reactions. *Clinical Psychology & Psychotherapy*, 29(2), 676–686. <https://doi.org/10.1002/cpp.2661>
- Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4), 450–469. <https://doi.org/10.1093/pan/mpq025>
- Vermunt, J. K., & Magidson, J. (2021). *Upgrade manual for latent GOLD basic, advanced/syntax, and choice version 6.0*. Statistical Innovations Inc.
- Williams, T. F., Vehabovic, N., & Simms, L. J. (2023). Developing and validating a facial emotion recognition task with graded intensity. *Assessment*, 30(3), 761–781. <https://doi.org/10.1177/10731911211068084>