TURUN YLIOPISTO

How observed mentalizing varies in narrating personally meaningful stories and how this variation relates to narcissistic traits

Psychology
Master's Thesis
Department of Psychology and Speech-Language Pathology

Author:
Marisa Anniliina Lehtisalo

5.5.2026
Helsinki

Master’s Thesis

Subject: Psychology
Author: Marisa Anniliina Lehtisalo
Title: How observed mentalizing varies in narrating personally meaningful stories and how this variation relates to narcissistic traits
Supervisor: Jarno Tuominen
Number of Pages: 42
Date: 6.6.2026

Mentalizing, a multidimensional and context-dependent capacity for reflecting on one's own and others' mental states, is known to break down in emotionally charged interpersonal moments. Narcissistic functioning, where self-image, validation, and social comparison are central, is one such context, but mentalizing in narcissism has mostly been studied with self-report measures. Clinical accounts also hold that grandiose and vulnerable features come together as narcissistic pathology becomes pronounced, yet how this combination shapes mentalizing in real interpersonal situations remains unclear. This study used continuous observational coding to examine how mentalizing varies in narratives of personally meaningful experiences, and how this variation relates to grandiose and vulnerable narcissistic traits. Participants (n = 74), high or low scorers on the NPI-13, recorded four narratives about being admired or socially excluded in youth and in the present. Two observers rated mentalizing continuously with a joystick, and participants completed the Reflective Functioning Questionnaire (RFQ). Self-reported and observed mentalizing were essentially uncorrelated. Across the whole sample, narrative context strongly shaped observed mentalizing: outcast and youth narratives elevated the probability of negative mentalizing, and the youth–present contrast in positive mentalizing was strongest for admiration narratives. Turning to narcissistic traits, vulnerable and grandiose narcissism showed no main effects, but they interacted in predicting the magnitude of negative mentalizing once it occurred. The findings suggest that current self-report-based research on mentalizing in narcissism may not reflect how mentalizing unfolds in interpersonal situations

Key Words: mentalizing, reflective functioning, grandiose narcissism, vulnerable narcissism, observational coding, personally meaningful narratives, social exclusion, self-report validity

Table of Contents

1. Introduction

People have always tried to make sense of their own and others' minds, and how well they manage to do so shapes nearly every aspect of social functioning, from intimate relationships to mental health and a stable sense of self (Bateman & Fonagy, 2016). This capacity, referred to as mentalizing, is not a fixed trait but a multidimensional and context-dependent process that is particularly likely to break down in emotionally charged interpersonal moments where it would matter most (Bateman & Fonagy, 2016; Luyten et al., 2024). One context in which the workings of mentalizing become especially pressing is narcissistic functioning, where self-image, validation, and social comparison are themselves at the center (Pincus & Lukowitsky, 2010). Yet what is currently known about narcissism and mentalizing comes almost entirely from self-report instruments, which may capture how individuals view their cognitive abilities rather than how those abilities are deployed in real interpersonal situations. The present thesis takes a complementary approach, examining mentalizing as it unfolds during the narration of personally meaningful experiences, and asking how this observed mentalizing relates to grandiose and vulnerable narcissistic traits, both separately and in combination.

1.1 Mentalizing

Mentalizing can be understood as the capacity to tolerate uncertainty about mental states, to avoid black-and-white thinking, and to distinguish one's own subjective experience from the reality experienced by others (Bateman & Fonagy, 2016). The term itself originated in French psychoanalytic writing (Fain & David, 1963), and a recent expert consensus selected mentalizing as the most generic umbrella term, defined as the ability to attribute mental states such as beliefs, intentions, emotions, and perceptions to oneself and to others (Quesque et al., 2024). Wendt and colleagues (2024), among others, refer to mentalizing as "mindreading ability", a relatively stable individual difference in how accurately a person can make inferences about other people's mental states.

This other-focused framing captures only part of the construct, however. Fonagy and Target (1997, as cited in Bateman & Fonagy, 2016) describe mentalizing as "the spine of our sense of self and identity". At its core, mentalizing is a second-order representational capacity: it allows a person to take their own mental states as an object of thought rather than simply being immersed in them.

Applied to the self, this dual stance means experiencing oneself both as a subject who acts and feels in the moment and as an object that can be reflected upon and, importantly, presented to others (Luyten et al., 2020). The capacities for reflecting on the self and on others have typically been seen as closely intertwined, since they are thought to draw on a shared neural core (Bateman & Fonagy, 2016). Açıl and colleagues (2025) provide recent empirical support for this dual structure: a shared brain pattern predicted both self- and other-mentalizing across independent samples, but the two could also be reliably told apart, with self-related thought engaging more anterior and medial regions and other-related thought more posterior and lateral ones.

Contemporary models conceptualize mentalizing as a multidimensional capacity organized around four polarities (automatic versus controlled, self-versus other, internal versus external, and cognitive versus affective), and imbalances between these polarities, rather than a global deficit, are thought to characterize different forms of psychopathology (Luyten et al., 2024). For this reason, mentalizing has been suggested as a transdiagnostic mechanism for understanding mental health problems more broadly (Luyten et al., 2024; Benzi et al., 2026). A recent network-analytic study illustrates this concretely: in a large sample of emerging adults, uncertainty about mental states functioned as a central bridge node connecting core personality vulnerabilities to maladaptive epistemic stances such as mistrust and credulity (Benzi et al., 2026).

The assessment of Mentalizing has become increasingly important in clinical practice, since mentalization plays a role in many mental health problems both as an explanatory factor and as a background vulnerability (Sharp, 2025; Benzi et al., 2026). Most empirical work in this field operates within the tradition developed by Bateman and Fonagy, in which mentalizing is conceptualised as the imaginative capacity to interpret behaviour in terms of mental states. This capacity is rooted in attachment, develops gradually through early caregiving, and tends to break down under attachment-related distress. Meta-analytic evidence supports this developmental view: across 23 studies and nearly 4,000 participants, more severe childhood maltreatment was associated with lower mentalizing capacity, with the strongest effects observed in child samples and in clinical populations (Yang & Huang, 2024). Difficulties in reflecting on one's own and others' mental states have been linked to emotion dysregulation, misinterpretations of others' intentions, impulsive behaviour, and unstable interpersonal relationships. For this reason, mentalization has been suggested as a transdiagnostic mechanism for understanding psychopathology more broadly (Luyten et al., 2024; Benzi et al., 2026).

Mentalizing also actively shapes behavior when it goes wrong. Bateman and Fonagy (2016) describe how mistakes in understanding oneself and others can lead individuals to act out in ways that are meant to retain mental stability and to attenuate feelings that have become incomprehensible. This is why mentalizing has become not just a diagnostic concept but also a treatment target. The same logic speaks to why mentalizing matters outside the clinic: if it organizes how a person makes sense of meaningful interpersonal events, then both stable individual differences and momentary fluctuations in this capacity should leave traces in the way personally meaningful experiences are described. Whether current assessment methods can capture this kind of variation is, however, far from clear.

1.2 Mentalization measures

From a clinical perspective, mentalization assessment offers something that traditional personality measures cannot reach. Personality measures can reliably tell us whether a person struggles with self-regulation or interpersonal relationships, but they don't tell us why these difficulties arise. A mentalization-based formulation tries to fill this gap by combining several complementary pieces of information: the person's overall mentalizing style (e.g. hyper- vs. hypomentalizing), the balance across the four mentalizing polarities (self vs. other, internal vs. external, cognitive vs. affective, and automatic vs. controlled), and the extent to which the person relies on non-mentalizing modes such as psychic equivalence (Sharp, 2025). How well current assessment tools can capture this multidimensional picture is, however, another question. As Sharp (2025) herself emphasizes, no single standardized measure covers all these aspects, and self-report instruments in particular reach only one slice of it.

Mentalization is most often assessed with self-report questionnaires or task-based performance measures, but the validity of both approaches has been increasingly questioned. The Reflective Function Scale (RFS), coded from Adult Attachment Interview narratives, has long been considered the gold standard, but its time-consuming administration and stringent training requirements have encouraged the development of more efficient alternatives (Stefana et al., 2024). These include self-report instruments such as the Reflective Functioning Questionnaire (RFQ), the Mentalization Questionnaire (MZQ), and the Mentalization Scale (MentS), as well as task-based measures like the Movie for the Assessment of Social Cognition (MASC) and the Reading the Mind in the Eyes Test (RMET) (Stefana et al., 2024; Wendt et al., 2024). Of these, the MASC is often regarded as one of the more ecologically valid task-based measures of mentalizing, as its video format situates mental-

state inference in interpersonally relevant scenarios (Lakhani et al., 2025). A more recent text-based alternative, the Mentalizing Vignettes Task (MVT-24), takes a similar approach using written vignettes and provides comparable information about a person's overall mentalizing style (e.g. hyper- vs. hypomentalizing) across different relational contexts (Lakhani et al., 2025; Sharp, 2025). The RMET, in contrast, has faced more criticism, as a recent scoping review found that the construct validity of its scores is rarely reported, and where reported, often inadequate (Higgins et al., 2024).

The most direct test of convergence between self-report and task-based measures comes from a preregistered multimethod study by Wendt and colleagues (2024), who showed that the two approaches are essentially unrelated at the latent level. Task-based performance tracked general cognitive ability, whereas self-reports seemed to reflect how individuals view their own mindreading rather than how accurately they infer mental states. A related integrative analysis by Asgarizadeh et al. (2025) found that confidence in one's own mindreading ability emerged as a distinct factor in self-report measures. Moreover, higher confidence was associated with poorer psychological functioning. This finding supports the view that mentalizing involves a capacity to tolerate uncertainty and maintain reflective humility, rather than certainty about the accuracy of one's interpretations. Sharp (2025) extends this critique by noting that even task-based measures such as the MASC and MVT-24 mainly capture the overall mentalization style and don't really reach the polarities or prementalizing modes outlined earlier. Some self-report instruments try to fill this gap. For example, MentS separates self- and other mentalizing (Dimitrijević et al., 2018), the Brief Mentalized Affectivity Scale (BMAS) targets the affective side of mentalizing (Greenberg et al., 2021), and the MZQ taps into psychic equivalence (Hausberg et al., 2012). However, no single tool currently covers the whole picture. Mentalization is also inherently dynamic and context-dependent, and breakdowns are most likely to happen in emotionally charged interpersonal situations, especially under stress and attachment activation.

These methodological concerns have led to explicit calls for a different research agenda. Luyten and colleagues (2024) argue that future studies should take a multidimensional approach to mentalizing, distinguish between state-like and trait-like aspects of the construct, and focus on problem-specific rather than only general mentalizing. The combination of brief self-report and observational coding used in the present study is in line with these recommendations. Observational coding of mentalizing in narrative material also has methodological precedent: Köber and colleagues (2019) showed that the standard reflective functioning rating scheme can be applied reliably to ordinary

life narratives, although mean scores tended to be lower than in clinical interviews because monologic narratives lack the targeted prompts that scaffold reflection. This suggests that personally meaningful narratives are a viable, if more demanding, context for observational mentalizing assessment.

For this reason, both Sharp (2025) and Wendt et al. (2024) emphasise the importance of moment-to-moment assessment of mentalization in real interaction. The present study addresses this gap by combining a brief self-report measure (RFQ) with observational continuous coding of mentalizing during the telling of personally meaningful narratives. The narrative task is well suited to this purpose, since recounting personally meaningful experiences is exactly the kind of emotionally charged situation in which moment-to-moment fluctuations in mentalizing are most likely to become visible (Sharp, 2025; Luyten et al., 2024). Personally meaningful narratives offer one practical route to capturing this kind of variation, and the rationale for using them is developed next.

1.3 Narcissism

From a psychodynamic perspective, pathological narcissism has been understood as developing from early relational disruptions: when caregivers fail to provide adequate mirroring, the child cannot integrate aggressive and envious feelings linked to disappointment, and instead develops a grandiose self-structure that defends against an underdeveloped, fragile core (Kernberg, 2004; Gabbard & Crisp, 2018). More recent dimensional models shift the focus from these developmental origins to how narcissism is expressed in adult interpersonal functioning (Cain et al., 2008; Pincus & Lukowitsky, 2010).

Within this dimensional tradition, two phenotypes are typically distinguished: grandiose narcissism, characterised by overt self-importance, extraversion, and status-seeking, and vulnerable narcissism, defined by hypersensitivity, shame, introversion, and an unstable sense of self (Pincus & Lukowitsky, 2010). At the grandiose end, the feeling of superiority remains relatively insulated from narcissistic injury, while in vulnerable narcissism the same defensive sense of being special is more easily threatened and is accompanied by elevated arousal in social interactions (Pincus & Lukowitsky, 2010). The two phenotypes are not assumed to occur in isolation: when narcissism reaches a pathological level, grandiose and vulnerable features tend to co-occur and individuals fluctuate between them, although one expression typically dominates in a given context (Gabbard & Crisp, 2018; Pincus & Lukowitsky, 2010).

The differential correlates of these phenotypes have begun to be mapped empirically. In a preregistered study with 956 participants, Tuominen et al. (in press) found that vulnerable narcissistic traits were associated with compromised self-reported mentalizing, elevated epistemic mistrust and credulity, a heightened need to belong, and pronounced internal and external shame, whereas grandiose narcissistic traits predicted lower belongingness needs, reduced shame, and epistemic mistrust without credulity or clear mentalizing deficits. This pattern fits with theoretical accounts in which vulnerable narcissism reflects an unstable self-experience contingent on social validation, while grandiose narcissism operates through more self-regulatory, defensive maintenance of a stable, though not necessarily accurate, self-image (Cain et al., 2008; Pincus & Lukowitsky, 2010).

1.4 Narcissism and mentalizing

Imbalances of mentalizing can go in two directions: too much certainty about mental states reflects hypermentalizing, while strong uncertainty reflects hypomentalizing (Bateman & Fonagy, 2016). When assessed through self-report, it is hypomentalizing rather than hypermentalizing that appears most relevant to narcissistic pathology. In a network analysis of pathological personality features, Benzi et al. (2026) found that uncertainty about mental states functioned as a structurally central feature, tightly linked both to the narcissistic vulnerability and self-impairment core and to maladaptive epistemic stances, while certainty about mental states remained peripheral. A similar asymmetry has been reported in the literature on childhood maltreatment: in community samples the link is specifically with hypomentalizing, whereas signs of hypermentalizing have been observed mainly in clinical groups (Yang & Huang, 2024).

For vulnerable narcissism, the picture from self-report studies is consistent across clinical and non-clinical samples. Blay et al. (2024) found that narcissistic vulnerability was negatively associated with self-mentalizing and overall mentalizing, even after adjusting for borderline and ADHD symptoms; the link between vulnerability and emotion dysregulation disappeared when self-mentalizing was added to the model, suggesting a mediating role. Tuominen et al. (in press) reported the same direction in a large non-clinical sample, where vulnerable traits were linked to both increased uncertainty and reduced certainty about mental states. Koskinen et al. (2025) further describe a self-defeating loop in which the external need for validation, poor mentalizing, epistemic mistrust, credulity, and shame feed into relational instability.

For grandiose narcissism, the same self-report evidence is less coherent. Blay et al. (2024) reported a small positive association with other-mentalizing, while the negative link with self-mentalizing did not survive adjustment for comorbidities. Tuominen et al. (in press) did not find a clear deficit either; if anything, grandiose traits came with greater confidence in interpreting mental states, which the authors interpret as a defensive or self-promotional stance rather than as reflective accuracy.

This pattern raises a measurement concern that underlies the remainder of this chapter. Mentalizing in narcissism has so far been assessed almost exclusively with self-report instruments, and it can be argued that these instruments capture how individuals view their own reflective abilities rather than how those abilities are deployed in real interpersonal situations. This concern seems especially relevant in the context of narcissism, where self-presentation is itself a defining feature, particularly in the grandiose form. The issue is revisited later in the chapter, after the rationale for using personally meaningful narratives has been developed.

1.5 Mentalizing and narratives: Admiration and exclusion

If mentalizing fluctuates substantially from moment to moment, then a single trait-level score will inevitably miss the very variation that matters most. Steinberg et al. (2024) made this point empirically by showing that about two thirds of the variance in observer-rated mentalizing was at the within-person level, meaning that the same individual can mentalize quite differently from one situation to the next. Personally meaningful narratives offer a useful alternative to questionnaires, because telling such a story brings the person closer to the original emotional state and to the relational context in which mentalizing was originally needed. Narratives can therefore be seen as a more ecologically valid window into mentalizing, since they activate the same kinds of emotional and interpersonal demands under which mentalizing tends to break down (Bateman & Fonagy, 2016; Wendt et al., 2024).

This dynamic view of mentalizing is particularly suited to studying narcissism. The features that are theoretically most central to narcissistic functioning, such as fluctuating self-esteem, sensitivity to admiration, and reactivity to perceived rejection, are unlikely to be captured by asking the person to describe themselves in general terms. Clinical and mentalization-based models conceptualise pathological narcissism not as a uniform deficit in understanding minds, but as a context-dependent imbalance that becomes visible specifically under interpersonal stress, when the grandiose self is activated or threatened.

Two narrative contexts were chosen for the present study: stories about being admired and stories about being socially excluded. These contexts are theoretically opposite poles of the same interpersonal dimension, and both carry strong relevance for self-esteem and the attachment system. Admiration narratives involve being seen, valued, and recognized by others, which typically supports a positive sense of self and a relatively secure relational stance. Exclusion narratives, on the other hand, involve being rejected, ignored, or left out, which activates threat to belonging, shame, and attachment distress. Both contexts are also emotionally engaging enough to evoke mentalizing in real time. The choice of these contexts also follows directly from the way narcissism has been conceptualized in recent work. Back et al. (2013) describe narcissistic functioning as organized around the pursuit of admiration and the defense against threats to the grandiose self, which means that situations involving recognition and situations involving rejection are precisely the ones in which narcissistic processes are expected to come to the surface.

Empirical work supports the relevance of these contexts for narcissism specifically. Fontana et al. (2026) found that adolescents produced less reflective narratives after experimentally induced social exclusion than after inclusion, and that this drop was most pronounced for those higher in pathological narcissism, even though the emotional distress of exclusion was shared across the sample. In other words, exclusion appeared to function as a context that selectively revealed mentalizing vulnerabilities tied to narcissism, rather than as a stressor that affected everyone's reflective capacity equally. The admiration and exclusion narratives are therefore not only two emotionally meaningful tasks, but also a way to bring the theoretical claims about narcissism, self-esteem regulation, and mentalizing into the same observable situation.

1.6 Narcissism observed: Self-presentation and the limits of self-report

Grandiose and vulnerable narcissism share underlying goals such as gaining esteem, status, and a sense of superiority, but they differ in the strategies they use to manage how others perceive them and in how confidently they manage to use them (Casale et al., 2016; Hart et al., 2017, 2019). Grandiose narcissism is associated with assertive self-presentation tactics that actively build a desired image: exaggerating accomplishments, claiming credit and special treatment, intimidating to gain influence, and disparaging rivals to seem superior by comparison (Hart et al., 2017, 2019). Cain, Pincus, and Ansell (2008) note that although the surface presentation looks like excessive self-confidence, it is often underpinned by markedly low self-esteem, which the individual attempts to protect by defensively denying awareness of negative aspects of the self.

Vulnerable narcissism shows a more layered self-presentation profile, drawing on both assertive and defensive tactics: presenting oneself as weak to gain sympathy, excuse-making, self-handicapping, and disclaimers, alongside attempts to claim positive identities (Hart et al., 2017, 2019). Casale et al. (2016) sharpen the contrast: grandiose narcissists actively promote a perfect image of themselves and feel pressured to appear effortlessly perfect, whereas vulnerable narcissists endorse every facet of perfectionistic self-presentation but perceive themselves as unable to project such an image. The clinical descriptions align with this picture, marked by shyness, social withdrawal, intensified monitoring of interpersonal interactions, and pervasive self-criticism (Pincus et al., 2014; Pincus & Lukowitsky, 2010). The two forms therefore differ not only in tactics but in the confidence with which those tactics can be deployed.

Psychophysiological work suggests that the confident image of the grandiose narcissist may not be as effortless as it appears. Koskinen et al. (2024, 2025) measured skin conductance and heart rate during natural conversations and found that participants high in grandiose narcissism showed elevated physiological arousal when talking about themselves, even though this arousal did not show up in their self-reports. Arousal was especially high when they told stories about being admired and rose further when the listener became disengaged (Koskinen et al., 2024). Vulnerable narcissism, by contrast, was associated with generally higher heart rate during narration, fitting with the broader picture of social interaction as more stressful in this group. What grandiose narcissists say they feel and what their bodies show, in other words, can be quite different things.

These behavioral and physiological findings point to a methodological concern that is directly relevant for the present study. If grandiose narcissists actively construct an image of effortless competence, and vulnerable narcissists closely monitor how they appear, self-report instruments are likely to be limited in what they can reveal about mentalizing. The concern is not, however, simply about deliberate distortion. Sleep et al. (2017) found that, in low-stakes research settings, neither grandiose nor vulnerable narcissism was associated with response invalidity, and vulnerable narcissism was negatively related to positive impression management. Simard et al. (2023) make a complementary point about empathy: current methods may underestimate true deficits in narcissistic individuals because the underlying capacity can be intact while the motivation to engage it is not. The limits of self-report in narcissism may therefore stem less from misrepresentation than from what self-reports can in principle access.

A direct illustration of this gap comes from Bilotta et al. (2018), who assessed mentalizing in patients with narcissistic personality disorder using both a self-report measure and a semi-structured

narrative interview based on a personally meaningful event. The same individuals appeared relatively spared on the self-report but markedly less reflective when their narrative accounts were rated by external observers. A similar mismatch emerges at the level of the broader social-cognition literature: in a systematic review, Eddy (2022) concluded that perspective-taking is not reliably impaired in narcissistic individuals, that affective forms of empathy are more often diminished, and that studies combining self-report with objective measures tend to find inflated self-evaluations relative to actual performance, particularly for grandiose features. Bateman and Fonagy (2016) provide a theoretical anchor: individuals with narcissistic features can produce talk that resembles mentalizing without necessarily reflecting the underlying capacity, and they may appear proficient in cognitive aspects of mentalizing while showing little sense of the emotional impact of their actions on others.

The present study contributes to this broader effort by examining mentalizing as it unfolds during the narration of personally meaningful events, asking whether the relatively intact mentalizing seen in grandiose narcissism on self-report measures is also observed when mentalizing is assessed from the outside.

1.7 The research question and hypotheses

While the link between narcissism and self-reported mentalizing is now reasonably well established, less is known about how narcissistic traits relate to mentalizing as it unfolds in personally meaningful stories. The literature reviewed above points to some gaps in the current research. Most previous studies have relied on self-report questionnaires, which limits our understanding of how narcissistic traits manifest in observed mentalizing. Moreover, even though narcissistic self-presentation has been studied before, the contrast between admiration and exclusion contexts, which is theoretically central to understanding how these traits operate in different social situations, has not yet been examined alongside observational mentalizing data.

This study aims to address these gaps by examining mentalizing as it is observed in narratives of personally meaningful experiences (admiration versus exclusion), and the extent to which grandiose and vulnerable narcissistic traits explain variation in it, both separately and in combination. The predictions developed below follow from the literature reviewed in the preceding sections.

Two predictions concern the narrative context. Given that mentalizing tends to break down under attachment-related distress and emotional arousal, and that exclusion has been shown to selectively

reveal mentalizing vulnerabilities (Bateman & Fonagy, 2016; Fontana et al., 2026), exclusion narratives are expected to challenge observed mentalizing more strongly than admiration narratives. A similar pattern is expected for narratives concerning past (youth) versus present events, since recalling earlier interpersonal experiences activates the same kinds of emotional and attachment-related demands.

Three predictions concern narcissistic traits. Drawing on the consistent self-report evidence linking vulnerable narcissistic features to compromised mentalizing (Blay et al., 2024; Tuominen et al., in press), vulnerable narcissism is expected to relate to lower observed mentalizing during personally meaningful stories. For grandiose narcissism, the picture is less straightforward: although self-report studies have not consistently shown grandiose deficits, observational and physiological evidence (Koskinen et al., 2024, 2025; Simard et al., 2023) suggests that the apparent intactness of mentalizing in grandiose narcissism may reflect self-presentation and confidence rather than genuine reflective capacity, leading to the prediction that the high grandiose narcissism group will show lower observed mentalizing than the low grandiose narcissism group. Finally, since pronounced narcissistic pathology tends to combine grandiose and vulnerable features rather than appear as either dimension in pure form (Gabbard & Crisp, 2018; Pincus & Lukowitsky, 2010), the negative association between vulnerable narcissism and observed mentalizing is expected to be amplified in the high grandiose narcissism group.

The following hypotheses were tested:

H1: Narrative conditions (story type and time) differ in observed mentalizing.

H2: Self-reported mentalizing ability (RFQ) is positively associated with observed mentalizing.

H3: Vulnerable narcissism is associated with lower mentalizing as observed in narrating personally meaningful stories.

H4: The high grandiose narcissism group (N+) shows lower mentalizing than the low grandiose narcissism group (N−).

H5: The association between vulnerable narcissism and observed lower mentalizing is increased in the high grandiose narcissism group (N+) compared to the low grandiose narcissism group (N−).

2. Methods

This study is part of a larger research project on narcissism, face, and social interaction at the University of Helsinki. The present focus is on the video-recorded narrative content from this broader project, and only the variables relevant to the current research questions are reported here.

2.1 Participants and procedure

Recruitment was carried out through Facebook advertising, social media, mailing lists, and University of Helsinki media channels, and it took place in two stages. First, all interested individuals filled in an online prescreening questionnaire that included demographic items (gender, age, education), the Narcissistic Personality Inventory-13 (NPI-13; Gentile et al., 2013), the vulnerable narcissism subscale of the Super-Brief Pathological Narcissism Inventory (SB-PNI; Schoenleber et al., 2015), and the Reflective Functioning Questionnaire (RFQ; Fonagy et al., 2016). Altogether 935 individuals completed this prescreening. Based on their NPI-13 scores, two gender- and age-matched groups were then invited to the laboratory phase: those scoring above 40 formed the high grandiose narcissism group (N+; n = 36), and those scoring below 27 formed the low-range control group (N−; n = 37). The cut-off values were taken from Henttonen et al. (2022) and corresponded approximately to the 80th and 30th percentiles of the NPI-13 distribution within the present prescreening sample. One participant in the N− group scored slightly above the lower cut-off (NPI-13 = 29) but was retained because they were the first individual to complete the experiment. The final laboratory sample consisted of 74 participants.

On the day of the experiment, each participant recorded four short narrative videos at Otaniemi following standardized instructions. The four narratives covered being admired in youth, being admired in the present, being socially excluded in youth, and being socially excluded in the present. Each video was approximately two minutes long.

Of the 74 participants, the majority identified as female (66.2%, n = 49), 31.1% as male (n = 23), and two participants (2.7%) did not report their gender. Participant ages ranged from 19 to 49 (M = 28.30, SD = 7.90). In terms of education, 44.6% reported upper secondary school as their highest degree (n = 33), 29.7% a bachelor's degree (n = 22), and 13.5% a master's degree (n = 10). Smaller groups reported vocational school (5.4%, n = 4), post-secondary vocational institute (2.7%, n = 2), comprehensive school (1.4%, n = 1), or a doctoral degree (1.4%, n = 1). One participant was excluded from the regression analyses due to a missing value on the vulnerable narcissism measure, leaving 73 participants for those analyses.

2.2 Ethics

The study received ethical approval from the Aalto University ethics board. Participation was entirely voluntary and participants were free to withdraw at any time without consequences. Those who additionally participated in the fMRI component of the larger research project received a €50 compensation for time and inconvenience. To qualify for participation, individuals had to be at least 18 years old and have no diagnosed neurological or psychiatric conditions (ICD diagnostic categories G and F). The study's privacy notice was made available to all participants (https://osf.io/ga6p9).

2.3 Self-report measures

Vulnerable narcissistic traits were assessed with the vulnerable narcissism subscale of the Super-Brief Pathological Narcissism Inventory (SB-PNI; Schoenleber et al., 2015). The SB-PNI is a 12-item self-report instrument with a 6-point Likert response format that taps both grandiose and vulnerable dimensions of pathological narcissism. It is a condensed form of the original 52-item Pathological Narcissism Inventory (Pincus et al., 2009), developed by retaining the best-performing items from each of the two original factors. In the present analyses, only the six items of the vulnerability subscale were used, with higher sum scores indicating more pronounced vulnerable narcissistic traits. Internal consistency in the present laboratory sample (n = 74) was good ($\alpha = .85$, $\omega = .85$).

Grandiose narcissistic traits were measured with the Narcissistic Personality Inventory-13 (NPI-13; Gentile et al., 2013). The NPI-13 contains 13 items rated on a 5-point Likert scale ranging from 1 (totally disagree) to 5 (totally agree). The scale can be decomposed into three subscales (Leadership/Authority, Grandiose Exhibitionism, and Entitlement/Exploitativeness), but in the present analyses I relied on the total sum score, where higher values reflect more pronounced grandiose narcissistic traits. Internal consistency in the present laboratory sample (n = 74) was excellent (α = .94, ω = .95). As noted above, NPI-13 scores were also the basis for assigning participants to the high (N+) and low (N−) grandiose narcissism groups.

Self-reported mentalizing was measured with the eight-item Reflective Functioning Questionnaire (RFQ; Fonagy et al., 2016), which targets two forms of disturbed mentalizing: hypomentalizing, characterised by excessive uncertainty about mental states, and hypermentalizing, characterised by excessive certainty about them. The RFQ can be scored either using the original two-factor solution that distinguishes Certainty and Uncertainty about mental states (Fonagy et al., 2016), or using the unidimensional approach proposed by Müller and colleagues (2022), where the two ends of a single continuum represent excessive certainty (low scores) and excessive uncertainty (high scores). In the present study, I used the unidimensional scoring approach, where the RFQ-Uncertainty (RFQ_U) and RFQ-Certainty (RFQ_C) subscales are combined into a single total RFQ score. Internal consistency of the total RFQ score in the present laboratory sample (n = 74) was good (α = .82, ω = .83).

Age, gender (male or female), and highest attained education level (classified as upper secondary or lower, BA, and MA or higher) were collected as background information.

2.4 Observed mentalizing: joystick method

Observed mentalizing was assessed with a continuous joystick coding method adapted from previous research and implemented in DARMA, a MATLAB-based software (sampling rate 20 Hz, bin size 0.5 s). Two psychology students rated each video independently. Before any actual rating took place, both raters were trained together using written coding guidelines. The aim of this training was to make sure that the raters had a shared and consistent understanding of what counts as positive and negative mentalizing.

Following the coding guidelines, positive mentalizing was defined as the narrator processing the situation and the minds of the people involved reflectively and flexibly, tolerating complexity and uncertainty about mental states, and remaining emotionally present in a way that fits the content of the story. Negative mentalizing was defined as the narrator describing situations one-dimensionally and inflexibly, expressing unwarranted certainty about their own or others' mental states, drawing overgeneralized conclusions from their own assumptions or behavior, and either failing to engage emotionally with content that calls for emotional involvement or being overwhelmed by strong affect.

During each video, the raters positioned the joystick continuously along a front-to-back axis to track these dimensions in real time, and raw joystick values were aggregated into 0.5-second bins, providing a temporal resolution of 500 ms. The lateral axis of the joystick was not used in the analyses; it was only included because the DARMA setup requires a second dimension for technical reasons. Each bin was then discretized into a three-point scale: +1 indicating positive mentalizing, 0 indicating a neutral state, and −1 indicating negative mentalizing. Ratings were time-stamped and aggregated into the percentage of time the joystick was in each position. Inter-rater agreement was moderate, with Cohen's $\kappa = .40$ across the full dataset and percentage agreement ranging from 76% to 78% across the four narrative conditions (admiration in youth = 78%, admiration in the present = 76%, exclusion in youth = 76%, exclusion in the present = 77%). A consensus scoring file was constructed in which each time point received a shared rating: when one rater assigned a neutral score (0) and the other assigned −1 or +1, the time point was coded as 0; time points where the two raters assigned opposing scores (+1 and −1) were treated as missing.

The present study examined all four narrative conditions: two admiration narratives (admiration in youth and admiration in the present) and two social exclusion narratives (exclusion in youth and exclusion in the present). The primary outcome variable was the percentage of time spent in negative mentalizing (percent_negative).

2.5 Analyses

All analyses were conducted in R using the packages glmmTMB, lme4, emmeans, and dplyr. Prior to modelling, the four narrative conditions were recoded into two crossed factors, StoryType (admiration vs. outcast) and Time (present vs. youth), and a dichotomous grandiose narcissism

variable, Group (N− vs. N+). Vulnerable narcissism (Vuln) and self-reported mentalizing (RFQ) were retained as continuous predictors and were mean-centred for use in the regression models.

The primary outcome variable, percentage of time spent in negative mentalizing (percent_negative), was strongly right-skewed and zero-inflated (Shapiro–Wilk $W = .518$, $p < .001$), which made standard linear regression inappropriate. We therefore analysed it with a two-part hurdle model: a logistic component predicting the occurrence of any negative mentalizing during a narrative segment (any_neg), and a Tweedie component (log link) modelling the magnitude of negative mentalizing among segments where percent_negative > 0. For each component, we built a sequence of nested models (M1–M5), beginning with the narrative context (StoryType × Time) and adding, step by step, vulnerable narcissism, grandiose narcissism group, the Vuln × Group interaction, and self-reported mentalizing. Models were compared by AIC and likelihood-ratio tests against the previous step. The same sequence was used to model positive mentalizing (percent_positive), which was less skewed but still well suited to a Tweedie family with a log link.

3. Results

Means, standard deviations, and Spearman intercorrelations between vulnerable and grandiose narcissism, self-reported and observed mentalizing, and the demographic covariates are presented in Table 1. Vulnerable narcissism (SB-PNI) had a mean of 17.18 (SD = 6.80) and grandiose narcissism (NPI-13) a mean of 34.11 (SD = 12.86). Self-reported mentalizing (RFQ total) had a mean of 8.47 (SD = 5.49), while the overall observed mentalizing mean across the four narratives was 0.13 (SD = 0.13).

Vulnerable and grandiose narcissism were strongly and positively correlated ($\rho = .74$, $p < .01$). This overlap is theoretically expected, since clinical accounts hold that pronounced narcissistic pathology tends to combine vulnerable and grandiose features rather than appear as one or the other in pure form (Pincus & Lukowitsky, 2010). However, it also has a methodological implication for the present analyses: the dichotomous N− vs. N+ grouping (based on NPI-13) is not independent of the continuous vulnerable narcissism score, and the two trait predictors should be interpreted as overlapping rather than as separate, additive sources of variance. We return to this point when interpreting the Vuln × Group interaction below. Neither dimension showed a strong association with self-reported or observed mentalizing: correlations between vulnerable narcissism and the

mentalizing variables were small in magnitude ($|\rho| \leq .22$), and the same was true for grandiose narcissism. Crucially, the observed mentalizing mean was essentially uncorrelated with self-reported mentalizing ($\rho = -.04$), already foreshadowing the formal test of H2 reported below.

Among the demographic covariates, age was positively associated with education ($\rho = .56$, $p < .01$), reflecting the expected tendency for older participants to have completed higher levels of formal education. Other associations of age and gender with the narcissism and mentalizing variables were small and mostly non-significant, suggesting that demographic differences were not a major driver of the psychological associations of interest in this sample.

Table 1. Descriptive statistics and Spearman intercorrelations between study variables

Variable	M	SD	Intercorrelations								
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1. Vulnerability	17.18	6.80	—								
2. Grandiosity	34.11	12.86	.74**	—							
3. Self-reported Mentalizing	8.47	5.49	−.22	−.09	—						
4. Negative Mentalizing mean	4.13	6.23	.17	.31**	−.00	—					
5. Positive Mentalizing mean	17.1	10.4	.02	−.07	−.10	−.20	—				
6. Mentalizing mean	0.13	0.13	−.09	−.22	−.04	−.52**	.90**	—			
7. Age	28.30	7.90	−.03	.03	.20	−.11	.04	.05	—		
8. Gender	—	—	.02	−.00	.11	−.08	−.22	−.14	−.06	—	
9. Education	—	—	−.17	−.03	.18	−.12	.05	.10	.56**	.11	—

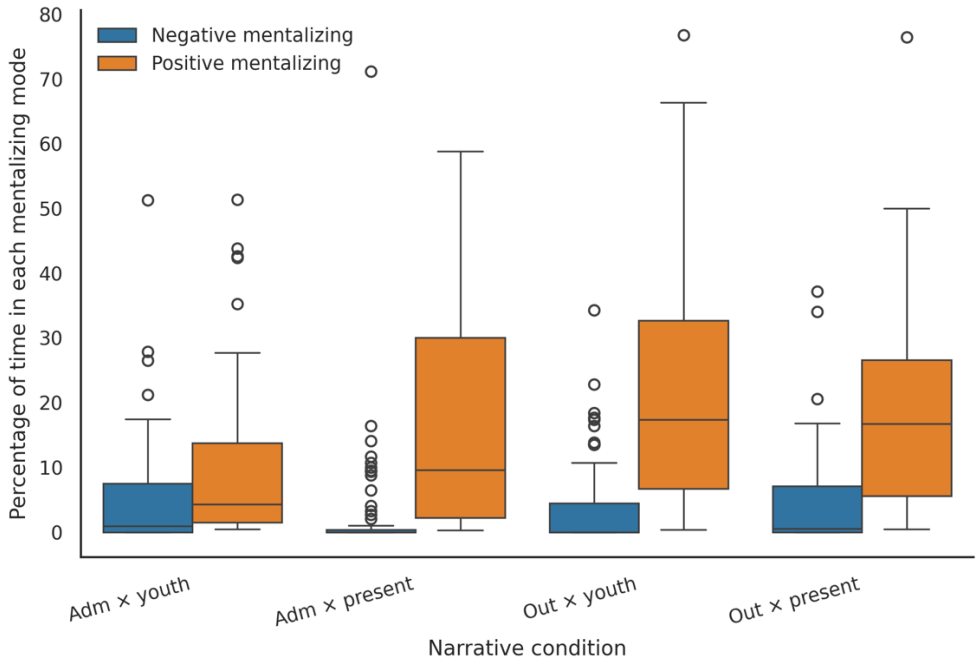
Note. Vulnerability = vulnerable narcissism (SB-PNI subscale); Grandiosity = Narcissistic Personality Inventory-13 total score; Self-reported Mentalizing = Reflective Functioning Questionnaire total score; Negative Mentalizing mean = the percentage of narrative time spent in negative mentalizing averaged across the four narrative conditions per participant; Positive Mentalizing mean = the percentage of narrative time spent in positive mentalizing averaged across the four narrative conditions per participant; Mentalizing mean = overall observed mentalizing quality averaged across the four narratives. Age is in years; Gender is coded 0 = male, 1 = female; Education is coded 0 = upper secondary or lower, 1 = lower university degree or higher. Correlations are Spearman rank-order. * $p < .05$, ** $p < .01$, *** $p < .001$.

Each participant contributed up to four narrative segments (admiration vs. outcast crossed with present vs. youth), yielding 292 segments with valid scores on vulnerable narcissism and the RFQ. Of these, valid joystick consensus codes for the percentage variables were available for 260

segments from 73 participants, which constituted the analysis dataset for the binary component of the hurdle model. Within those 260 segments, negative mentalizing (percent_negative > 0) was observed in 115 segments across 59 participants, and the Tweedie component was fit on this subset. Positive mentalizing (percent_positive > 0) was observed in all 260 segments, indicating that some positive mentalizing was present in nearly every narrative. Negative mentalizing, by contrast, was substantially sparser and more variable across segments, a pattern that motivated the hurdle approach rather than a single linear model for the negative outcome.

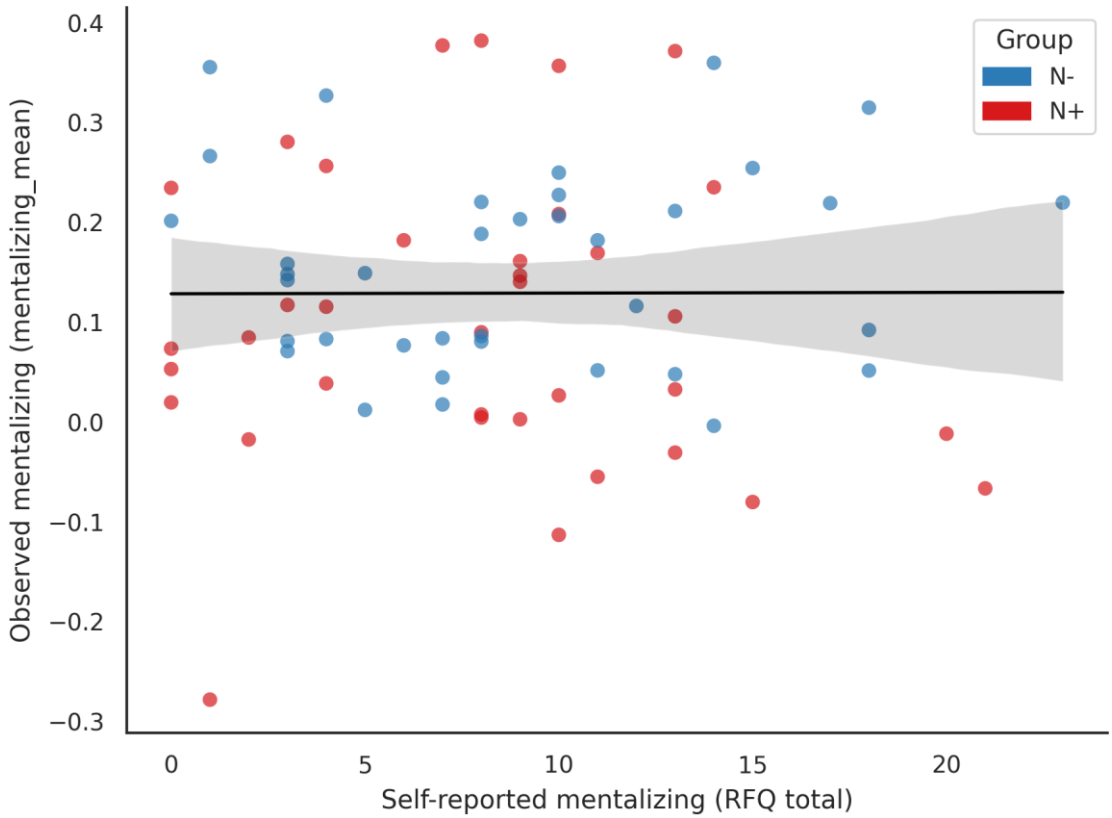
Figure 1 displays box plots of the percentage of time spent in positive and in negative mentalizing for each of the four narrative conditions (Admiration × Present, Admiration × Youth, Outcast × Present, Outcast × Youth). Across all conditions, the distribution of positive mentalizing lay clearly above that of negative mentalizing, indicating that participants generally spent more time in positive than in negative mentalizing modes during their narratives. At the same time, the size of the boxes and whiskers within each condition pointed to substantial individual variability, suggesting that the narrative manipulation operated against a background of meaningful between-person differences.

Figure 1. Box plots of positive and negative mentalizing across narrative conditions



To test H2, we computed a Spearman correlation between self-reported mentalizing (RFQ total) and the mean observed mentalizing score, using one observed mentalizing value per participant. The association was essentially zero ($\rho = -.009$, $p = .939$), providing no evidence that participants who reported better mentalizing on the RFQ also displayed more effective mentalizing during the narrative task. Figure 2 illustrates this result with a scatterplot of RFQ scores against the observed mentalizing mean, with a fitted linear regression line. The cloud of points was widely dispersed along the RFQ axis but showed no systematic trend in relation to observed mentalizing, visually confirming the near-zero correlation. In short, self-perceived mentalizing ability did not predict how much participants actually mentalized when narrating their personally meaningful stories.

Figure 2. Scatterplot of self-reported mentalizing (RFQ total) against the observed mentalizing mean



3.1 Hurdle models for negative mentalizing (H3–H5)

To test the hypotheses concerning narcissistic traits and observed mentalizing (H3–H5), we fit a sequence of nested models for each component of the hurdle. The model comparisons are

summarised in Table 2 and the coefficient estimates from the most extensive model (M5) are reported in Table 3. Because the model sequence corresponds to the pre-specified set of hypotheses (narrative context → vulnerable narcissism → grandiose group → their interaction → self-reported mentalizing), we treated the comparisons as confirmatory and did not apply a multiple-comparison correction across the steps. The description below focuses on the substantive findings rather than on each step of the sequence.

In the binary component, the narrative manipulation produced robust and large effects: the odds of any negative mentalizing were strongly elevated in outcast narratives ($\beta = 1.514$, $SE = 0.432$, $z = 3.50$, $p < .001$) and in youth narratives ($\beta = 1.436$, $SE = 0.434$, $z = 3.31$, $p < .001$), and the StoryType × Time interaction was also reliable ($\beta = -1.819$, $SE = 0.597$, $z = -3.05$, $p = .002$), reflecting that the additive effects of telling an outcast story and recalling the youth period did not simply combine but moderated each other. By contrast, none of the trait predictors meaningfully changed the probability of negative mentalizing once narrative context was accounted for: vulnerable narcissism ($\beta = 0.020$, $p = .658$), grandiose narcissism group ($\beta = 0.122$, $p = .806$), the Vuln × Group interaction ($\beta = 0.041$, $p = .589$), and RFQ ($\beta = -0.016$, $p = .645$) were all far from significance, and the corresponding model comparisons (M2–M5) showed no improvement in fit, although the addition of vulnerable narcissism alone in M2 approached significance ($\chi^2 = 3.06$, $p = .080$).

The Tweedie component, which models the magnitude of negative mentalizing among segments where it was present, showed a different pattern. None of the narrative-context terms reached significance (StoryType $\beta = -0.028$, $p = .921$; Time $\beta = 0.192$, $p = .474$; StoryType × Time $\beta = -0.084$, $p = .809$), and the main effects of vulnerable narcissism ($\beta = -0.035$, $p = .152$), grandiose group ($\beta = 0.353$, $p = .208$), and RFQ ($\beta = 0.004$, $p = .834$) were also non-significant. Adding the Vuln × Group interaction in M4 produced the only meaningful improvement in fit over the entire sequence ($\Delta AIC = -3.31$; $\chi^2 = 5.31$, $p = .021$); the interaction term itself was significant in M5 ($\beta = 0.096$, $SE = 0.040$, $z = 2.40$, $p = .017$), and adding RFQ in M5 left fit essentially unchanged ($\Delta AIC = +1.96$, $p = .834$). We therefore treat M4 (without RFQ) as the primary model for testing the trait-level hypotheses, with M5 reported in Table 3 to document the (null) contribution of self-reported mentalizing.

Table 2. Sequential model comparisons for negative mentalizing

Component / Model	Predictor added	AIC	ΔAIC	χ^2	df	p
Binary component **Any negative mentalizing**						
M1	StoryType × Time	344.93	—	—	—	—
M2	+ Vuln_c	343.87	−1.06	3.06	1	.080
M3	+ Group	345.76	+1.89	0.11	1	.737
M4	+ Vuln_c × Group	347.52	+1.76	0.24	1	.623
M5	+ RFQ_c	349.30	+1.78	0.21	1	.645
Tweedie component **Magnitude of negative mentalizing**						
M1	StoryType × Time	735.73	—	—	—	—
M2	+ Vuln_c	735.14	−0.59	2.59	1	.107
M3	+ Group	733.67	−1.47	3.47	1	.063
M4	+ Vuln_c × Group	730.36	−3.31	5.31	1	**.021***
M5	+ RFQ_c	732.32	+1.96	0.04	1	.834

Note. Vuln_c = mean-centred vulnerable narcissism; Group = high vs. low grandiose narcissism (N+ vs. N−); RFQ_c = mean-centred RFQ total score. ΔAIC = change in AIC relative to the previous model. χ^2 and p are likelihood ratio test statistics from anova(), comparing each model to the previous one. The binary component uses the full dataset (N rows = 260, 73 participants); the Tweedie component uses only observations with percent_negative > 0 (N rows = 115, 59 participants). * p < .05.

Table 3. Coefficient estimates from the final hurdle model (M5)

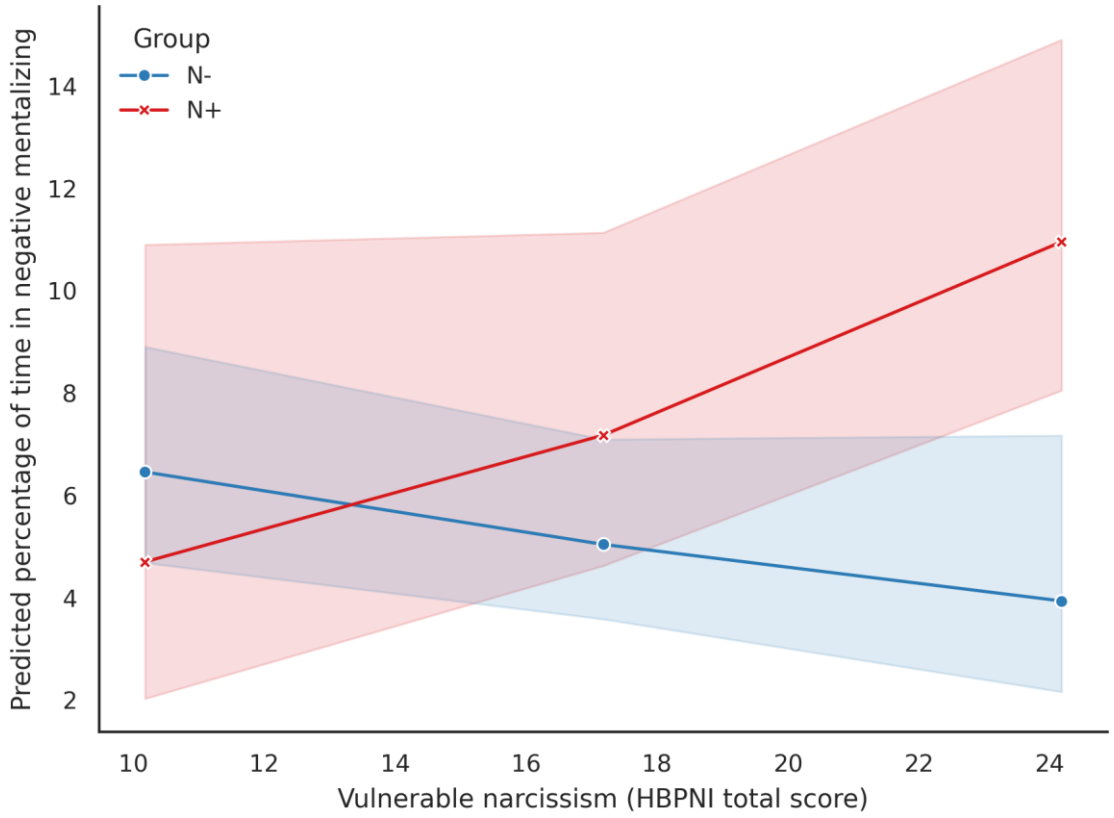
Predictor	β	SE	z	p	β	SE	z	p
	Binary component				**Tweedie component**			
Narrative context								
StoryType (Outcast)	1.514	0.432	3.502	**<.001**	−0.028	0.280	−0.100	.921
Time (Youth)	1.436	0.434	3.309	**<.001**	0.192	0.268	0.716	.474
StoryType × Time	−1.819	0.597	−3.045	**.002**	−0.084	0.348	−0.241	.809

Narcissistic traits								
Vuln_c	0.020	0.045	0.443	.658	−0.035	0.025	−1.432	.152
Group (N+ vs. N−)	0.122	0.496	0.245	.806	0.353	0.280	1.259	.208
RFQ_c	−0.016	0.034	−0.461	.645	0.004	0.019	0.209	.834
Interaction								
Vuln_c × Group	0.041	0.076	0.541	.589	0.096	0.040	2.396	**.017**

Note. Estimates are from Model M5 of the hurdle analysis. The binary component (left side) is a logistic regression predicting whether any negative mentalizing occurred (any_neg). The Tweedie component (right side) models the magnitude of negative mentalizing (percent_negative) among observations where percent_negative > 0. StoryType is coded 0 = admiration, 1 = outcast; Time is coded 0 = present, 1 = youth; Group is coded 0 = N−, 1 = N+; Vuln_c and RFQ_c are mean-centred vulnerable narcissism (SB-PNI) and RFQ total scores. M4 (without RFQ_c) is treated as the primary model for H3–H5, and adding RFQ_c in M5 did not improve model fit in either component (ΔAIC ≈ +1.8 in the binary part; ΔAIC ≈ +2.0 in the Tweedie part).

To clarify the form of the Vuln × Group interaction, we computed simple slopes of vulnerable narcissism within each combination of StoryType and Group from the M5 Tweedie model (see Appendix). In admiration narratives, the slopes were positive in both groups but did not reliably differ from zero (N−: $\beta = 0.061$, 95% CI [−0.057, 0.179]; N+: $\beta = 0.116$, 95% CI [−0.006, 0.238]). In outcast narratives, the picture was more nuanced: the slope was small and negative in the low-grandiosity group, just excluding zero (N−: $\beta = -0.052$, 95% CI [−0.101, −0.002]), and small and positive but not reliable in the high-grandiosity group (N+: $\beta = 0.047$, 95% CI [−0.019, 0.113]). The interaction therefore reflects a divergence of slopes between the two groups rather than a strong directional effect within either group, and the most reliable individual simple slope in the model is in fact a small negative one for the low-grandiosity group during outcast stories. Figure 3 visualises the predicted percentages of negative mentalizing across the range of vulnerable narcissism for the two groups, and the divergence of slopes is more visible there than the absolute height of either line.

Figure 3. Predicted negative mentalizing across vulnerable narcissism by grandiose narcissism group



Two considerations qualify the interaction. First, the Tweedie component is fit on 115 segments from 59 participants, which after the further split between N− and N+ leaves roughly 30 participants per cell as the effective basis for the interaction. The Wald confidence intervals on the simple slopes reflect this: they are wide, and their boundaries with zero are close. The result should therefore be regarded as a relatively fragile pattern that requires replication. Second, given the strong correlation between vulnerable and grandiose narcissism (ρ = .74), the Group factor is not independent of the continuous Vuln predictor: high-grandiosity participants tend also to score relatively highly on vulnerable narcissism. The Vuln × Group interaction can therefore be read either as a true crossover between two distinct trait dimensions or, more parsimoniously, as a non-linear effect of overall narcissistic pathology that is amplified in participants who are high on both dimensions. We treat the latter reading as the more conservative one and discuss its implications below.

In exploratory analyses (see Appendix), we extended the M4 model with three-way interactions involving the narrative-context factors. Adding StoryType × Vuln × Group approached but did not

reach conventional significance in either component (binary: $\chi^2 = 7.70$, $p = .053$; Tweedie: $\chi^2 = 6.27$, $p = .099$), suggesting that the Vuln × Group pattern may differ between admiration and outcast narratives, although the present sample is underpowered to test this reliably. The corresponding Time × Vuln × Group interaction was not significant in either component (binary $\chi^2 = 2.24$, $p = .525$; Tweedie $\chi^2 = 2.92$, $p = .404$). These analyses were not part of the pre-specified hypothesis sequence and are reported here as flags for future research rather than as substantive findings.

3.2 Positive mentalizing

Sequential model comparisons for positive mentalizing are summarised in Table 4. The baseline model M1, which included only the StoryType × Time interaction, already provided a good description of the data (AIC = 1976.24). Adding vulnerable narcissism in M2 (ΔAIC = +2.00, $p = .931$), grandiose group in M3 (ΔAIC = +0.59, $p = .234$), the Vuln × Group interaction in M4 (ΔAIC = +0.82, $p = .277$), and RFQ in M5 (ΔAIC = −0.10, $p = .147$) did not yield meaningful improvements, and none of the corresponding likelihood-ratio tests was significant. The narrative manipulation, rather than narcissistic traits or self-reported mentalizing, accounted for most of the systematic variation in positive mentalizing.

Table 4. Sequential model comparisons for positive mentalizing

Model	Predictors added	AIC	ΔAIC	χ^2	df	p
M1_pos	StoryType × Time	1976.24	—	—	—	—
M2_pos	+ Vuln_c	1978.24	+2.00	0.01	1	.931
M3_pos	+ Group	1978.82	+0.59	1.41	1	.234
M4_pos	+ Vuln_c × Group	1979.64	+0.82	1.18	1	.277
M5_pos	+ RFQ_c	1979.54	−0.10	2.10	1	.147

Note. Outcome is percent_positive (percentage of time in positive mentalizing across narratives). Vuln_c = mean-centred vulnerable narcissism; Group = high vs. low grandiose narcissism (N+ vs. N−); RFQ_c = mean-centred RFQ total score. ΔAIC = change in AIC relative to the previous model. χ^2 and p are likelihood ratio test statistics from anova() comparing each model to the previous one.

Coefficient estimates from M5_pos are presented in Table 5. The intercept reflects the predicted percentage of time spent in positive mentalizing for present-time admiration stories at the mean level of vulnerable narcissism and RFQ in the low-grandiosity group, and indicates that these

narratives were, on average, strongly dominated by positive mentalizing ($\beta = 2.80$, $SE = 0.20$, $z = 14.32$, $p < .001$). Consistent with the model-comparison results, the only reliable predictors were Time and the StoryType × Time interaction: youth stories showed significantly lower positive mentalizing than present-time stories ($\beta = -0.62$, $SE = 0.19$, $z = -3.22$, $p = .001$), and this Time effect was further moderated by StoryType ($\beta = 0.76$, $SE = 0.27$, $z = 2.82$, $p = .005$). The Time effect was clearly negative in admiration narratives, whereas in outcast narratives the difference between youth and present-time stories was smaller and not reliably different from zero on the log scale. In other words, telling youth admiration stories tended to pull participants out of positive mentalizing more strongly than telling youth outcast stories, relative to their present-time counterparts.

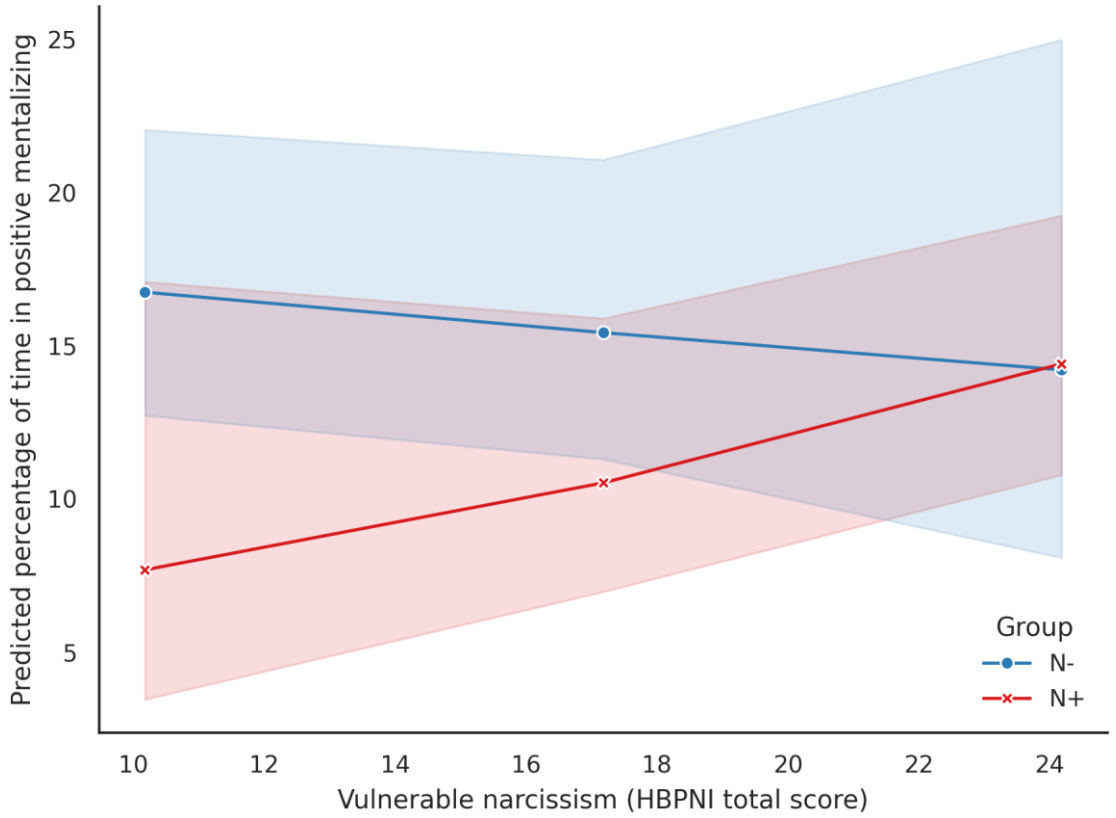
In contrast, the coefficients for vulnerable narcissism ($\beta = -0.01$, $p = .611$), grandiose group ($\beta = -0.38$, $p = .139$), RFQ ($\beta = -0.03$, $p = .151$), and the Vuln × Group interaction ($\beta = 0.06$, $p = .157$) were all small and non-significant, indicating that individual differences in narcissistic traits and self-reported mentalizing did not systematically predict the magnitude of positive mentalizing once narrative context was accounted for. Figure 4 plots the predicted percentage of time in positive mentalizing across the vulnerable narcissism range for the N− and N+ groups. Consistent with the non-significant Vuln × Group term, the two lines were nearly parallel and showed only minor changes across the vulnerability range. Taken together, the modelling results suggest that positive mentalizing was driven primarily by the temporal framing and content of the stories, while negative mentalizing, once it occurred, was sensitive to the combination of vulnerable and grandiose narcissistic traits.

Table 5. Tweedie coefficients from the final positive mentalizing model (M5_pos)

Predictor	**β**	**SE**	**z**	**p**
(Intercept)	2.80	0.20	14.32	**< .001***
StoryType (out vs. adm)	0.09	0.19	0.50	.621
Time (youth vs. present)	−0.62	0.19	−3.22	**.001***
Vuln_c	−0.01	0.02	−0.51	.611
Group (N+ vs. N−)	−0.38	0.26	−1.48	.139
RFQ_c	−0.03	0.02	−1.44	.151
StoryType × Time	0.76	0.27	2.82	**.005***
Vuln_c × Group	0.06	0.04	1.42	.157

Note. Estimates are from m5_pos_tw (Tweedie family, log link). β coefficients are on the log scale; positive values indicate higher predicted percent_positive. StoryType is coded 0 = admiration, 1 = outcast; Time is coded 0 = present, 1 = youth; Group is coded 0 = N−, 1 = N+. * p < .01.

Figure 4. Predicted positive mentalizing across vulnerable narcissism by grandiose narcissism group



The Vuln × Group interaction reported in the main analyses was further explored by adding three-way interactions involving the narrative-context factors. Two model comparisons were conducted, one for StoryType × Vuln × Group and one for Time × Vuln × Group, with both components of the hurdle (binary and Tweedie) examined separately. Table 5 summarizes the StoryType × Vuln × Group comparison, and Table 7 the Time × Vuln × Group comparison. Neither three-way interaction produced a meaningful improvement in fit: adding StoryType × Vuln × Group did not improve the binary component (ΔAIC = +4.23, $\chi^2 = 1.77$, $p = .622$), and adding Time × Vuln × Group improved fit in neither component (binary ΔAIC = +3.75, $\chi^2 = 2.25$, $p = .522$; Tweedie ΔAIC

= +3.25, $\chi^2 = 2.76$, $p = .431$). Some values for the Tweedie StoryType × Vuln × Group comparison were not produced by the model fit and are reported as missing in Table 5.

Table 5. StoryType × Vulnerable narcissism × Group interaction for negative mentalizing

Component	Model	Predictors added	AIC	ΔAIC	χ^2	*p*
Binary: any_neg	Baseline	Time + StoryType + Vuln_c × Group	355.61	–	–	–
	StoryType × Vuln_c × Group + Time	StoryType × Vuln_c × Group terms	359.84	+4.23	1.77	.622
Tweedie: percent_negative	Baseline	Time + StoryType + Vuln_c × Group	—	–	–	–
	StoryType × Vuln_c × Group + Time	StoryType × Vuln_c × Group terms	732.30	—	—	—

Note. The binary model comparison did not improve fit when the StoryType × Vuln_c × Group terms were added. The Tweedie comparison output did not provide the baseline model's AIC or the likelihood-ratio test statistics, so those values are left blank rather than inferred.

To clarify whether the simple slopes of vulnerable narcissism within each combination of narrative context and Group followed a consistent pattern, EM trends were computed from the Tweedie component (Tables 6 and 8). Splitting by StoryType and Group (Table 6) showed slightly more reliable slopes than the omnibus interaction. The slope for the low-grandiosity group during outcast narratives was small and negative, with a confidence interval just excluding zero ($\beta = -0.064$, 95% CI [−0.125, −0.002]); the remaining cells gave slopes whose confidence intervals comfortably included zero. Splitting by Time and Group (Table 8) produced one cell where the confidence interval also just excluded zero, namely the high-grandiosity group during present-time narratives ($\beta = 0.079$, 95% CI [0.004, 0.154]); the other three cells were non-significant. These individual cells should be interpreted with care, since the model comparison did not support either three-way interaction and the cells contain modest numbers of participants. They are reported as patterns to be examined in larger samples rather than as confirmatory evidence.

Table 6. Vulnerable narcissism slopes by StoryType and Group in the Tweedie model for negative mentalizing

StoryType	Group	Vuln_c trend	*SE*	95% CI lower	95% CI upper
Admiration	N−	−0.00447	0.0325	−0.0682	0.0593
Admiration	N+	0.06218	0.0447	−0.0254	0.1498

Outcast	N−	−0.06392	0.0314	−0.1254	−0.00245
Outcast	N+	0.05665	0.0381	−0.0179	0.1312

Note. Slopes are averaged over Time.

Table 7. Time × Vulnerable narcissism × Group interaction for negative mentalizing

Component	**Model**	**Predictors added**	**AIC**	**ΔAIC**	**χ^2**	***p***
Binary: any_neg	Baseline	StoryType + Time + Vuln_c × Group	355.61	–	–	–
	Time × Vuln_c × Group + StoryType	Time × Vuln_c × Group terms	359.36	+3.75	2.25	.522
Tweedie: percent_negative	Baseline	StoryType + Time + Vuln_c × Group	728.41	–	–	–
	Time × Vuln_c × Group + StoryType	Time × Vuln_c × Group terms	731.66	+3.25	2.76	.431

Note. Adding the Time × Vuln_c × Group terms did not improve model fit in either the binary or Tweedie component. The baseline model is the same for both components, with StoryType, Time, and the Vuln_c × Group interaction included.

Table 8. Vulnerable narcissism slopes by Time and Group in the negative mentalizing model

Time	**Group**	**Vuln_c trend**	***SE***	**95% CI lower**	**95% CI upper**
Present	N−	−0.0259	0.0314	−0.0875	0.0356
Present	N+	0.0790	0.0381	0.0043	0.1536
Youth	N−	−0.0439	0.0316	−0.1058	0.0180
Youth	N+	0.0265	0.0462	−0.0640	0.1171

Note. Slopes are averaged over StoryType. The positive slope for N+ in the present condition suggests a possible increase in negative mentalizing with higher vulnerable narcissism in that subgroup, but the overall three-way interaction was not supported by the model comparison.

4. Discussion

The aim of the present study was to examine how observed mentalizing varies in narratives of personally meaningful stories, and to what extent grandiose and vulnerable narcissistic traits explain variation in it. Three findings stand out. First, self-reported mentalizing on the RFQ was essentially uncorrelated with observed mentalizing during the narrative task, both at the bivariate level and as a predictor in the regression models. Second, narrative context exerted a strong and reliable influence on the occurrence of negative mentalizing and on the magnitude of positive mentalizing, while leaving the magnitude of negative mentalizing largely untouched. Third, narcissistic traits showed

no main effects on observed mentalizing once narrative context was accounted for, but the interaction between vulnerable narcissism and the high grandiose narcissism group did predict the magnitude of negative mentalizing in segments where negative mentalizing occurred. Together these findings suggest that what people say about their own reflective abilities is not the same thing as how their reflective abilities unfold in interpersonally meaningful situations, and that narcissistic traits become visible in observed mentalizing mainly when grandiose and vulnerable features combine, in line with clinical accounts of narcissistic pathology as a configuration rather than as separate dimensions.

Relevant single finding was the near-zero correlation between self-reported and observed mentalizing ($\rho = -.009$). H2 had predicted a positive association on the assumption that the RFQ should track at least some of the variance in how participants reflectively engaged with their own and others' mental states during the narratives. The data offered no support for this prediction, and adding self-reported mentalizing to either component of the hurdle model left model fit essentially unchanged. This pattern aligns closely with Wendt et al. (2024), who showed that self-report and task-based measures of mentalizing were unrelated at the latent level, and with Asgarizadeh et al. (2025), who found that self-reported confidence in mindreading emerged as a distinct factor that was, if anything, negatively related to psychological functioning. Sharp (2025) has argued that self-report instruments mainly index the person's image of their own reflective abilities, and the present results provide further empirical support for this view in a different observational paradigm. They also matter for the interpretation of the narcissism findings discussed below: if RFQ scores reflect self-presentation as much as actual reflective capacity, then the absence of a self-reported mentalizing deficit in grandiose narcissism (Blay et al., 2024; Tuominen et al., in press) cannot be taken as evidence that the underlying capacity is intact.

Narrative context emerged as the dominant source of variation in observed mentalizing, in line with H1. Outcast and youth narratives were both substantially more likely than admiration and present-time narratives to elicit any negative mentalizing, with large effects in the binary component of the hurdle model. The story type by time interaction in the same component indicated that the additive effect of telling an outcast story and recalling youth did not simply combine; the contrast was sharpest when one of the two demands was added to a present-time admiration baseline. For positive mentalizing, the strongest effect was instead a time by story type interaction in which youth admiration stories pulled participants further out of positive mentalizing than youth outcast stories, relative to their present counterparts. This pattern is consistent with the broader claim that

mentalizing fluctuates with attachment- and emotion-related demands rather than functioning as a stable trait (Bateman & Fonagy, 2016; Steinberg et al., 2024), and it converges with Fontana et al. (2026), who showed that exclusion-related contexts selectively reveal mentalizing vulnerabilities. Importantly, the magnitude of negative mentalizing once it occurred was not reliably affected by narrative context, suggesting that the situational pull operates more strongly on whether breakdowns happen at all than on how severe they become.

H3 and H4 were not supported. Neither vulnerable narcissism nor the grandiose group contrast produced reliable main effects on observed mentalizing in either component of the hurdle model, although the directional estimates in the Tweedie component pointed toward more negative mentalizing at higher levels of the trait. This null pattern stands in contrast to the self-report literature linking vulnerable traits to reduced mentalizing (Blay et al., 2024; Tuominen et al., in press). It tempers, but does not contradict, the prediction in H4 that grandiose narcissism may mask underlying mentalizing deficits behind a confident self-presentation.

H5 received partial support, restricted to the Tweedie component of the negative mentalizing model. The interaction between vulnerable narcissism and the high versus low grandiose narcissism group reached significance, but the simple-slope decomposition was thinner than the omnibus interaction suggested. The only reliable simple slope was a small negative association in the low-grandiosity group during outcast narratives (95% CI [−0.125, −0.002]). The corresponding slope in the high-grandiosity group was non-significantly positive, and the remaining cells comfortably included zero. The interaction therefore reflects a divergence of slopes between the two groups rather than a reliable directional effect within either group. The most reliable individual slope is the unexpected negative one in the low-grandiosity group, a pattern that warrants its own theoretical attention. Why higher vulnerable narcissism is associated with less negative mentalizing in this group during exclusion narratives is not predicted by current models, and may reflect something specific to how vulnerable features operate when they are not accompanied by overt grandiosity. The interaction is nevertheless broadly consistent with clinical accounts in which pronounced narcissistic pathology combines grandiose and vulnerable features rather than appearing as either dimension in pure form (Gabbard & Crisp, 2018; Pincus & Lukowitsky, 2010), and with the configuration described by Koskinen et al. (2025) as a self-defeating loop of validation-seeking, poor mentalizing, and shame. However, given the strong correlation between vulnerable and grandiose narcissism in this sample ($\rho = .74$), the same interaction can be read more parsimoniously as a non-linear effect of overall narcissistic pathology, amplified in participants who score highly on

both dimensions. The two readings are not mutually exclusive, and the present design cannot adjudicate between them.

The findings support a layered picture of mentalizing in narcissism that is consistent with mentalization-based theory but only partially with the self-report literature. What looks intact when participants describe their reflective abilities does not necessarily translate into what is observed when they narrate emotionally meaningful events, particularly under the kind of interpersonal stress that exclusion narratives and youth recollections appear to introduce. At the same time, observed mentalizing was not uniformly worse for participants with more pronounced narcissistic features, which suggests that the relationship is conditional rather than linear and that narcissistic patterns become most visible in observed mentalizing where vulnerable and grandiose features co-occur. This is in line with Bateman and Fonagy's (2016) caution that individuals with narcissistic features can produce talk that resembles mentalizing without necessarily reflecting the underlying capacity, and it supports the methodological argument made by Sharp (2025) and Wendt et al. (2024) for combining self-report instruments with observational coding rather than relying on either alone.

4.1.Limitations

Several limitations qualify the conclusions that can be drawn from these results. The most fundamental concerns statistical power. The laboratory sample comprised 73 participants, and the Tweedie component of the hurdle model, on which the H5 interaction was tested, was fit on 115 segments from 59 participants. Once split between the high and low grandiose narcissism groups, this leaves roughly 30 participants per cell as the effective basis for the interaction, and the Wald confidence intervals around the simple slopes reflect this thinness: their boundaries with zero are close, and the result should be regarded as a fragile pattern that requires replication before stronger claims are made about its direction.

A second concern follows from the strong correlation between vulnerable and grandiose narcissism in this sample ($\rho = .74$). The pre-registered design treated these as analytically separable predictors, with vulnerable narcissism as a continuous score and grandiose narcissism as a dichotomous group based on NPI-13 scores, but in practice the grandiose narcissism group factor is not independent of the continuous vulnerable narcissism score. As noted in the discussion of the interaction between vulnerable narcissism and the grandiose narcissism group, this overlap means that the trait-level results cannot fully distinguish between two distinct narcissism dimensions and a single underlying

pathology axis on which both dimensions load. The recruitment design, which selected participants from the tails of the NPI-13 distribution and then measured vulnerable narcissism continuously within those groups, accentuates rather than reduces this confound.

The observational coding method introduces its own limitations. Inter-rater agreement was moderate at best (Cohen's $\kappa = .40$, with percentage agreement of 76–78% across the four narrative conditions), and the consensus rule that treated opposing rater scores as missing reduced the effective number of valid time points. Although the joystick approach has methodological precedent in continuous-coding work and follows the spirit of Köber et al. (2019), it is not yet standardized for mentalizing assessment, and the present rating guidelines placed positive and negative mentalizing on a single front-to-back axis in a way that may have constrained the dimensionality of what raters could capture. Mean scores in narrative material also tend to be lower than in clinical interviews (Köber et al., 2019), as monologic narratives lack the targeted prompts that scaffold reflection in instruments such as the Reflective Functioning Scale.

The sample is non-clinical and largely Finnish, predominantly female (66.2%), and concentrated in the lower adult age range ($M = 28.30$ years), which limits generalizability to clinical narcissism and to populations outside the Nordic university recruitment context. One participant in the low grandiose narcissism group scored slightly above the lower NPI-13 cut-off (29 against a threshold of 27) and was retained for procedural reasons; this is unlikely to have meaningfully affected the group-level results but is a small departure from the pre-specified inclusion criteria. Finally, the narrative task itself involves self-presentation. Although Sleep et al. (2017) found no general response invalidity in narcissistic individuals in low-stakes settings, narratives recorded on video at a research site are not low-stakes in quite the same sense as anonymous self-report questionnaires, and the present design cannot fully isolate observed reflective capacity from how participants chose to present themselves on camera.

Despite these limitations, the present design has features that strengthen the inferences within its scope. The combination of brief self-report and observational coding follows the multimethod recommendations made by Luyten et al. (2024) and Wendt et al. (2024); the four within-subject narrative conditions allow narrative context to be modelled rather than treated as a confound; and the hurdle approach makes appropriate distributional assumptions about a zero-inflated outcome that would otherwise have violated the assumptions of standard linear modelling.

4.2 Recommendations for future research

The most direct priority for future work is replication of the interaction between vulnerable narcissism and the grandiose narcissism group in a substantially larger sample. Adequately powered tests of the three-way interactions that approached but did not reach significance in the present exploratory analyses (story type by vulnerable narcissism by grandiose narcissism group, $p = .053$ in the binary component) would also clarify whether this interaction pattern is specific to outcast narratives, as the simple-slope decomposition tentatively suggested, or generalizes across narrative contexts.

A second priority is to address the trait-level confound rather than treat it as a nuisance. Designs that recruit pure grandiose, pure vulnerable, and combined high-narcissism groups, rather than selecting on grandiosity alone and measuring vulnerability continuously within those groups, would allow a sharper test of whether the observed mentalizing patterns reflect distinct dimensions or a single overall pathology axis. Clinical samples with diagnosed narcissistic personality disorder would extend the inferences beyond the dimensional, non-clinical range covered here and connect the present paradigm to the clinical literature in which mentalization-based formulations are most directly applied (Sharp, 2025).

Methodological development of the observational coding is a third priority. Higher inter-rater agreement is achievable with more extensive joint training, a third arbitrating rater, and possibly automated supplementation of the joystick traces from speech and behavioral features. Decoupling positive and negative mentalizing onto separate axes, rather than placing them on opposite ends of a single front-to-back continuum, would also let raters capture configurations in which both modes are present within the same segment, which the current setup forces into a single rating.
Beyond the present design, the joystick paradigm could be combined with the kinds of physiological measurement reported by Koskinen et al. (2024, 2025), who showed that grandiose narcissists exhibit elevated physiological arousal during self-narration that is not visible in self-reports. Whether the same dissociation between self-report and behavior appears at the level of mentalizing, that is, whether grandiose narcissists who report intact reflective capacity also show physiological signs of effortful self-monitoring during narration, is a directly testable question that follows from the present results. Finally, the choice of narrative contexts could be extended. Admiration and exclusion were selected because they map onto the theoretically central poles of

narcissistic self-regulation, but other emotionally meaningful contexts, for instance shame, betrayal, or pride, would test whether the patterns observed here are specific to these two contexts or generalize across narratives that activate different aspects of self-experience.

References

Açıl, D., Andrews-Hanna, J. R., Lopez-Sola, M., van Buuren, M., Krabbendam, L., Zhang, L., van der Meer, L., Fuentes-Claramonte, P., Pomarol-Clotet, E., Salvador, R., Debbané, M., Vrticka, P., Vuilleumier, P., Sbarra, D. A., Coppola, A. M., White, L. O., Wager, T. D., & Koban, L. (2025). Brain neuromarkers predict self- and other-related mentalizing across adult, clinical, and developmental samples. *bioRxiv*. https://doi.org/10.1101/2025.03.10.642438

Asgarizadeh, A., Shoumali, R., & Tahan, M. (2025). The common structure of mentalizing. *PLOS ONE*, *20*(9), e0332722. https://doi.org/10.1371/journal.pone.0332722

Back, M. D., Küfner, A. C. P., Dufner, M., Gerlach, T. M., Rauthmann, J. F., & Denissen, J. J. A. (2013). Narcissistic admiration and rivalry: Disentangling the bright and dark sides of narcissism. *Journal of Personality and Social Psychology*, *105*(6), 1013–1037. https://doi.org/10.1037/a0034431

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

Benzi, I. M. A., Fontana, A., Carone, N., Locati, F., Parolin, L., & Ensink, K. (2026). Pathology of the self: A network analysis of personality functioning, narcissistic vulnerability, mentalizing, and epistemic trust across trauma profiles. *Borderline Personality Disorder and Emotion Dysregulation*, *13*, 7. https://doi.org/10.1186/s40479-026-00335-5

Bilotta, E., Carcione, A., Fera, T., Moroni, F., Nicolò, G., Pedone, R., Pellecchia, G., Semerari, A., & Colle, L. (2018). Symptom severity and mindreading in narcissistic personality disorder. *PLoS ONE*, *13*(8), e0201216. https://doi.org/10.1371/journal.pone.0201216

Blay, M., Bouteloup, M., Duarte, M., Hasler, R., Pham, E., Nicastro, R., Jan, M., Debbané, M., & Perroud, N. (2024). Association between pathological narcissism and emotion regulation: The role of self-mentalizing? *Personality and Mental Health*, *18*(3), 227–237. https://doi.org/10.1002/pmh.1613

Cain, N. M., Pincus, A. L., & Ansell, E. B. (2008). Narcissism at the crossroads: Phenotypic description of pathological narcissism across clinical theory, social/personality psychology,

and psychiatric diagnosis. *Clinical Psychology Review*, *28*(4), 638–656. https://doi.org/10.1016/j.cpr.2007.09.006

Casale, S., Fioravanti, G., Rugai, L., Flett, G. L., & Hewitt, P. L. (2016). The interpersonal expression of perfectionism among grandiose and vulnerable narcissists: Perfectionistic self-presentation, effortless perfection, and the ability to seem perfect. *Personality and Individual Differences*, *99*, 320–324. https://doi.org/10.1016/j.paid.2016.05.026

Dimitrijević, A., Hanak, N., Altaras Dimitrijević, A., & Jolić Marjanović, Z. (2018). The Mentalization Scale (MentS): A self-report measure for the assessment of mentalizing capacity. *Journal of Personality Assessment*, *100*(3), 268–280. https://doi.org/10.1080/00223891.2017.1310730

Eddy, C. M. (2022). Self-serving social strategies: A systematic review of social cognition in narcissism. *Current Psychology*, *41*(7), 4574–4597. https://doi.org/10.1007/s12144-021-01661-3

Fain, M., & David, C. (1963). Aspects fonctionnels de la vie onirique. *Revue Française de Psychanalyse*, *27*, 241–343.

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

Fonagy, P., & Target, M. (1997b). Research on intensive psychotherapy with children and adolescents. *Child and Adolescent Psychiatric Clinics of North America*, *6*, 39–51.

Fontana, A., Sideli, L., Cianfanelli, B., Somma, A., & Fossati, A. (2026). Navigating peer inclusion and exclusion: Pathological narcissism, reflective functioning, and rejection-related emotions in adolescents. *Clinical Neuropsychiatry*, *23*(1), 90–99. https://doi.org/10.36131/cnfioritieditore20260109

Gabbard, G. O., & Crisp, H. (2018). *Narcissism and its discontents: Diagnostic dilemmas and treatment strategies with narcissistic patients*. American Psychiatric Association Publishing.

Gentile, B., Miller, J. D., Hoffman, B. J., Reidy, D. E., Zeichner, A., & Campbell, W. K. (2013). A test of two brief measures of grandiose narcissism: The Narcissistic Personality Inventory–13 and the Narcissistic Personality Inventory–16. *Psychological Assessment*, *25*(4), 1120–1136. https://doi.org/10.1037/a0033192

Greenberg, D. M., Rudenstine, S., Alaluf, R., & Jurist, E. L. (2021). Development and validation of the Brief-Mentalized Affectivity Scale: Evidence from cross-sectional online data and an urban community-based mental health clinic. *Journal of Clinical Psychology*, *77*(11), 2638–2652. https://doi.org/10.1002/jclp.23203

Hart, W., Adams, J., Burton, K. A., & Tortoriello, G. K. (2017). Narcissism and self-presentation: Profiling grandiose and vulnerable narcissists' self-presentation tactic use. *Personality and Individual Differences*, *104*, 48–57. https://doi.org/10.1016/j.paid.2016.06.062

Hart, W., Richardson, K., Tortoriello, G. K., & Breeden, C. J. (2019). Revisiting profiles and profile comparisons of grandiose and vulnerable narcissism on self-presentation tactic use. *Personality and Individual Differences*, *151*, 109523. https://doi.org/10.1016/j.paid.2019.109523

Hausberg, M. C., Schulz, H., Piegler, T., Happach, C. G., Klöpper, M., Brütt, A. L., Sammet, I., & Andreas, S. (2012). Is a self-rated instrument appropriate to assess mentalization in patients with mental disorders? Development and first validation of the Mentalization Questionnaire (MZQ). *Psychotherapy Research*, *22*(6), 699–709. https://doi.org/10.1080/10503307.2012.709325

Henttonen, P., Salmi, J., Peräkylä, A., & Krusemark, E. A. (2022). Grandiosity, vulnerability, and narcissistic fluctuation: Examining reliability, measurement invariance, and construct validity of four brief narcissism measures. *Frontiers in Psychology*, *13*, 993663. https://doi.org/10.3389/fpsyg.2022.993663

Higgins, W. C., Kaplan, D. M., Deschrijver, E., & Ross, R. M. (2024). Construct validity evidence reporting practices for the Reading the Mind in the Eyes Test: A systematic scoping review. *Clinical Psychology Review*, *108*, 102378. https://doi.org/10.1016/j.cpr.2023.102378

Kernberg, O. F. (2004). *Aggressivity, narcissism, and self-destructiveness in the psychotherapeutic relationship: New developments in the psychopathology and psychotherapy of severe personality disorders*. Yale University Press.

Köber, C., Kuhn, M. M., Peters, I., & Habermas, T. (2019). Mentalizing oneself: Detecting reflective functioning in life narratives. *Attachment & Human Development*, *21*(4), 313–331. https://doi.org/10.1080/14616734.2018.1473886

Koskinen, E., Henttonen, P., Harjunen, V., Krusemark, E., Piispanen, M., Voutilainen, L., Wuolio, M., & Peräkylä, A. (2024). Putting self at stake by telling a story: Storyteller's narcissistic

traits modulate physiological emotional reactions to recipient's disengagement. *PLoS ONE*, *19*(8), e0302703. https://doi.org/10.1371/journal.pone.0302703

Koskinen, E., Henttonen, P., Kettunen, S. K., Pesonen, S., Piispanen, M., Voutilainen, L., Wuolio, M., & Peräkylä, A. (2024). Shame in social interaction: Descriptions of experiences of shame by participants with high or low levels of narcissistic traits. *British Journal of Social Psychology*, *63*(3), 1429–1449. https://doi.org/10.1111/bjso.12734

Koskinen, E., Henttonen, P., Harjunen, V., Krusemark, E., Salmi, J., Tuominen, J., Wuolio, M., & Peräkylä, A. (2025). 'Wired up about self': Narcissistic traits predict elevated physiological arousal during self-disclosure in conversation. *International Journal of Psychophysiology*, *210*, 112527. https://doi.org/10.1016/j.ijpsycho.2025.112527

Lakhani, S., Bhola, P., Mehta, U. M., & Bhaskarapillai, B. (2025). Development and preliminary validation of the Mentalizing Vignettes Task: A measure of mentalizing across relational contexts. *Journal of Personality Assessment*, *107*(6), 743–755. https://doi.org/10.1080/00223891.2025.2509501

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

Müller, S., Wendt, L. P., & Zimmermann, J. (2022). Development and validation of the Certainty About Mental States Questionnaire (CAMSQ): A self-report measure of mentalizing oneself and others. *Assessment*, *30*(3), 651–674. https://doi.org/10.1177/10731911211061280

Pincus, A. L., Ansell, E. B., Pimentel, C. A., Cain, N. M., Wright, A. G. C., & Levy, K. N. (2009). Initial construction and validation of the Pathological Narcissism Inventory. *Psychological Assessment*, *21*(3), 365–379. https://doi.org/10.1037/a0016530

Pincus, A. L., Cain, N. M., & Wright, A. G. C. (2014). Narcissistic grandiosity and narcissistic vulnerability in psychotherapy. *Personality Disorders: Theory, Research, and Treatment*, *5*(4), 439–443. https://doi.org/10.1037/per0000031

Pincus, A. L., & Lukowitsky, M. R. (2010). Pathological narcissism and narcissistic personality disorder. *Annual Review of Clinical Psychology*, *6*, 421–446. https://doi.org/10.1146/annurev.clinpsy.121208.131215

Quesque, F., Apperly, I., Baillargeon, R., Baron-Cohen, S., Becchio, C., Bekkering, H., Bernstein, D., Bertoux, M., Bird, G., Bukowski, H., Call, J., Catmur, C., Cheng, Y., Conway, J. R.,

Decety, J., Dimaggio, G., Doherty, M., Ferrari, P. F., Fonagy, P., … Brass, M. (2024). Defining key concepts for mental state attribution. *Communications Psychology*, *2*, 29. https://doi.org/10.1038/s44271-024-00077-6

Schoenleber, M., Roche, M. J., Wetzel, E., Pincus, A. L., & Roberts, B. W. (2015). Development of a brief version of the Pathological Narcissism Inventory. *Psychological Assessment*, *27*(4), 1520–1526. https://doi.org/10.1037/pas0000158

Sharp, C. (2025). Comment on: Development and preliminary validation of the Mentalizing Vignettes Task: A measure of mentalizing across relational contexts. *Journal of Personality Assessment*, *107*(6), 807–808. https://doi.org/10.1080/00223891.2025.2563886

Simard, P., Simard, V., Laverdière, O., & Descôteaux, J. (2023). The relationship between narcissism and empathy: A meta-analytic review. *Journal of Research in Personality*, *102*, 104329. https://doi.org/10.1016/j.jrp.2022.104329

Sleep, C. E., Sellbom, M., Campbell, W. K., & Miller, J. D. (2017). Narcissism and response validity: Do individuals with narcissistic features underreport psychopathology? *Psychological Assessment*, *29*(8), 1059–1064. https://doi.org/10.1037/pas0000413

Stefana, A., Jolić Marjanović, Z., & Dimitrijević, A. (2024). The brief version of the Mentalization Scale (MentS-12): Evidence-based assessment of mentalizing capacity. *Journal of Personality Assessment*, *106*(6), 740–749. https://doi.org/10.1080/00223891.2024.2326884

Steinberg, N., Moshe-Cohen, R., Sheena, L., Lifshitz, M., Berger, R., & Levy-Gigi, E. (2024). Capturing a mentalized moment: A pilot study of the psychometric properties of a novel assessment method of mentalizing in daily life. *Current Psychology*, *43*, 7596–7611. https://doi.org/10.1007/s12144-023-04963-w

Tuominen, J., Henttonen, P., Piispanen, M., Koskinen, E., Wuolio, M., Harjunen, V. J., Krusemark, E., Peräkylä, A., & Salmitaival, J. (in press). To be known is to be exposed: How vulnerable and grandiose narcissistic traits differentially affect shame, mentalizing, belongingness, and epistemic trust. *Journal of Personality Disorders*. https://doi.org/10.31234/osf.io/83uk9_v3

Wendt, L. P., Zimmermann, J., Spitzer, C., & Müller, S. (2024). Mindreading measures misread? A multimethod investigation into the validity of self-report and task-based approaches. *Psychological Assessment*, *36*(5), 365–378. https://doi.org/10.1037/pas0001310

Yang, L., & Huang, M. (2024). Childhood maltreatment and mentalizing capacity: A meta-analysis. *Child Abuse & Neglect*, *149*, 106623. https://doi.org/10.1016/j.chiabu.2023.106623

Appendices

Appendix 1: Statement on the use of generative AI

Claude, a Large Language Model developed by Anthropic, was used during the process of writing this paper. More precisely, this tool assisted in writing and reviewing code used for the data analysis. All outputs were reviewed and verified by the author.