BMJ Open

The validity and reliability of self-report measures of mentalising for adults: a protocol for systematic review

Ahmad Asgarizadeh , Amir Hossein Daneshmand Kafteroudi, Mohammad Ali Mazaheri, Saeed Ghanbari

To cite: Asgarizadeh A, Daneshmand Kafteroudi AH, Mazaheri MA, *et al*. The validity and reliability of self-report measures of mentalising for adults: a protocol for systematic review. *BMJ Open* 2025;**15**:e107520. doi:10.1136/bmjopen-2025-107520

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2025-107520).

Received 06 July 2025
Accepted 02 September 2025

Education and Psychology Faculty, Shahid Beheshti University, Tehran, Iran

Correspondence to
Professor Mohammad Ali Mazaheri;
mazaheri45@gmail.com

ABSTRACT

Introduction Mentalising, the capacity to understand behaviour via underlying mental states, is a key construct in psychopathology. While self-report instruments are widely used to assess mentalising, significant questions about their psychometric properties persist and no systematic review has comprehensively evaluated them using standardised criteria. This systematic review, guided by the COnsensus-based Standards for the selection of health Measurement Instruments () methodology, aims to: (1) identify all available self-report mentalising measures for adults; (2) evaluate the methodological quality of their validation studies; (3) synthesise and grade the quality of evidence on their measurement properties and (4) provide evidence-based recommendations for their use in research and clinical practice.

Methods and analysis Five electronic databases (SCOPUS, Web of Science, PsycINFO, PubMed, ProQuest) will be searched from their inception, supplemented by a search of grey literature and reference lists. We will include studies of any design that report on at least one measurement property of a self-report measure of mentalising in adults. Two reviewers will independently screen all records, extract relevant data and assess the methodological quality of included studies using the COSMIN Risk of Bias checklist. For each instrument, the evidence for each measurement property will be synthesised, and the overall quality of the evidence will be graded using a modified Grading of Recommendations Assessment, Development and Evaluation approach.

Ethics and dissemination As this systematic review will synthesise data from previously published studies, it does not require formal ethical approval. The findings will be disseminated through a peer-reviewed, open-access publication and presentations at scientific conferences. The results will provide a comprehensive inventory of available measures and a rigorous evaluation of their psychometric quality, creating an evidence base to guide clinicians and researchers in selecting the most appropriate instruments for mentalising assessment.

PROSPERO registration number CRD420251031469.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ This preregistered protocol, which adheres to the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015 statement, applies the standardised COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) methodology for systematic reviews of patient-reported outcome measures.

⇒ The review process incorporates multiple features to enhance methodological rigour, including a comprehensive search strategy developed with an information specialist, independent dual review for all stages and an independent assessment of each measure's content validity by the review team.

⇒ A key methodological challenge will be applying the COSMIN Risk of Bias checklist to primary studies that were likely not designed with these specific standards in mind, which may impact the quality assessment.

⇒ The conceptual heterogeneity of mentalising across different instruments and over time may pose challenges for the consistent application of content validity criteria and direct comparison between measures.

⇒ The review's scope is limited to English-language publications and excludes measures validated only in highly specific populations, which may affect the comprehensiveness and generalisability of the findings.

INTRODUCTION

The ability to interpret behaviour through the lens of underlying mental states, termed mentalising, represents an essential human capacity for navigating social life. Contemporary frameworks conceptualise mentalising across several intersecting dimensions: automatic/implicit versus controlled/explicit processing, where automatic involves fast, reflexive responses requiring little effort while controlled reflects deliberate thinking demanding awareness; self-oriented versus other-oriented understanding, distinguishing between mentalising one's own mental states versus those of others; cognitive versus affective components, separating the ability to recognise and reason about mental states from the capacity to understand their emotional significance and resonate with them and focus on internal versus external

features, differentiating between inferences based on inner knowledge versus observable cues like facial expressions.[1]

When the ability to mentalise breaks down, individuals may revert to more primitive modes of psychological functioning. One such mode is *psychic equivalence*, where the line between one's mind and the outside world blurs, leading to an unshakeable conviction that internal feelings are objective facts. This is exemplified when a traumatic memory is not just recalled but is intensely re-experienced in the present moment. Another mode is the *teleological stance*, where the validity of intentions or emotions is judged solely by tangible, physical results. This creates a focus on demonstrable acts over verbal assurances, meaning a feeling like love might only seem real if proven with a physical gesture. Finally, in *pretend mode*, the inner world becomes uncoupled from any real-world grounding, leading to a form of intellectualisation known as 'pseudomentalising'. Individuals can speak extensively about their feelings, yet the conversation remains superficial and lacks authentic emotional weight.[2]

Disruptions in mentalising capacity have been linked to various forms of psychopathology. Several meta-analyses and narrative reviews provide compelling evidence that most forms of psychological disorders are associated with mentalising deficits, which serve as important indicators of functional impairment.[3 4] Specific associations have been documented with personality disorders,[5 6] anxiety and related disorders,[7] eating disorders[8] and developmental and psychotic disorders.[9]

The validity of these findings, however, depends critically on reliable and valid measurement of mentalising capacity. A variety of methodologies have emerged to assess this construct, each with inherent limitations in fully capturing its complexity. The Reflective Functioning Scale (RFS),[10] considered the gold standard, requires extensive training and involves time-consuming administration/scoring procedures, limiting its feasibility for both clinical and research purposes. Additionally, there is still a lack of evidence for some of its measurement properties (eg, sensitivity to change and test-retest reliability).[11]

Task-based measures present different challenges. Although numerous task-based measures are developed to assess general social cognition, to the best of our knowledge, no such measure has been developed to assess mentalising capacity. Notably, the task-based social cognition measures are commonly administered to samples with profound mentalising deficits (eg, patients with schizophrenia or autism spectrum disorder) and may not be applicable to non-clinical samples due to ceiling effects.[12 13] Conceptually, these measures also tend to emphasise cognitive/other-oriented aspects while neglecting affective/self-oriented components essential in clinical contexts. More concerning is recent evidence suggesting task performance correlates strongly with general cognitive ability (r=0.85), calling into question whether these tasks measure mentalising specifically or general cognitive capacity.[14]

Given these limitations, self-report instruments have emerged as practical alternatives, offering accessibility and cost-effectiveness for large-scale studies. Among the most widely employed are the Reflective Functioning Questionnaire (RFQ),[15] the Mentalization Questionnaire (MZQ)[16] and the Mentalization Scale (MentS).[17] Each of these instruments, while making valuable contributions to the field, presents potential limitations that warrant closer examination.

The eight-item RFQ was designed to assess certainty and uncertainty about mental states, providing an efficient screening measure for research and clinical settings. However, questions have been raised regarding its conceptual foundation, psychometric structure and scoring approach. Recent studies with clinical and non-clinical samples have suggested potential issues with its dimensionality, balance across mentalising domains and discriminant validity from related constructs.[18] Thus, the RFQ may not adequately capture the theoretical complexity of mentalising across its multiple dimensions, particularly in its assessment of other-oriented processes.

The 15-item MZQ has provided a valuable tool for assessing the affective dimensions of mentalising. This measure appears to emphasise self-related mentalising aspects while providing a more limited assessment of other-oriented processes, raising doubts about its comprehensive coverage of mentalising. Additionally, substantial shared variance ($r \approx 0.60$) has been identified between the MZQ and emotion dysregulation measures,[19 20] suggesting potential limitations in its discriminant validity.

Similarly, the 28-item MentS represents an important contribution by assessing three components (self-related mentalising, other-related mentalising and motivation to mentalise), potentially offering a more balanced approach than other instruments. However, some concerns may be noted about this measure, including the neglect of the automatic/controlled dimension, its inclusion of items about general psychological interest and the positive associations between its other-related dimension and narcissistic features.[21] Particularly, this positive correlation directly contradicts theoretical expectations and raises fundamental questions about what the measure is actually assessing.

A recent study by Wendt *et al*[14] on other-oriented mentalising (termed 'mindreading ability') raises yet another important question about self-report assessment approaches. Their findings suggest that self-reported mindreading might primarily reflect perceived competence (a 'mindreading self-concept') rather than the actual capacity. When controlling for positive self-evaluations, self-reported mindreading scores demonstrated negative associations with psychosocial functioning. This suggests that current instruments may measure confidence in other-oriented mentalising abilities rather than their actual sophistication.

Another particularly significant concern involves the operationalisation of prementalising modes. Despite their importance in clinical conceptualisation and therapeutic

techniques, empirical validation of these modes through dedicated measurement remains limited. As Duschinsky and Foster[22] note, 'We do not actually know empirically that the three modes of non-mentalising are negatively associated with mentalising, or how strongly; We do not know that the reduction in non-mentalising is associated with reduced symptoms or greater quality of life…' (p. 288). Furthermore, contemporary theory emphasises the importance of contextual flexibility in mentalising, highlighting how mentalising capacities may vary across different relationships and situations.[23 24] Our review will thus examine whether and how existing measures address this contextual component.

Despite these significant measurement challenges, no systematic review has comprehensively evaluated the validity and reliability of self-report mentalising measures using standardised assessment criteria. This gap is particularly concerning as measurement method significantly influences research outcomes: studies have demonstrated stronger associations between mentalising and psychopathology when both constructs are measured via self-report, compared to when task-based measures of mentalisation are used.[25 26] Such method effects, rather than true construct relationships, may lead to erroneous conclusions about the role of mentalising in psychopathology.

This systematic review aims to address this critical gap by applying the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN)[27 28] methodology to assess the psychometric properties of existing self-report mentalising measures. Specifically, this review aims to:

1. Identify available self-report measures that assess mentalising in adults.
2. Evaluate the methodological quality of studies examining the psychometric properties of these measures using the COSMIN methodology.
3. Classify and grade the quality of evidence presented in research papers.
4. Summarise the evidence on the measurement properties of each identified measure.
5. Provide recommendations regarding the use of the most psychometrically sound measures.

METHODS AND ANALYSIS

The review protocol was preregistered with PROSPERO (CRD420251031469). This systematic review protocol is reported in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P)[29] and COSMIN guidelines[28] to ensure comprehensive and transparent reporting. This process commenced in April 2025 and is anticipated to be completed by January 2026.

Eligibility criteria

Studies will be selected for inclusion based on predefined criteria for the population, instruments, comparators, outcomes and study design. A detailed summary of the inclusion and exclusion criteria is provided in table 1.

Information sources

We will search the following electronic databases from inception to the search date: SCOPUS, Web of Science Core Collection (including Science Citation Index and Social Science Citation Index), PsycINFO (via EBSCO), PubMed and ProQuest Dissertations & Theses Global. These databases were selected to ensure comprehensive coverage of psychology, psychiatry and health sciences literature. Grey literature will be searched through ProQuest Dissertations & Theses Global. Additionally, we will hand-search reference lists of included studies and relevant systematic reviews to identify additional eligible studies. No date restrictions will be applied to maximise comprehensiveness.

Search strategy

We collaboratively devised the search strategies based on existing guidelines, using terms alternative to mentalising (eg, 'mentalisation', 'reflective functioning') and terms related to measurement properties (eg, 'psychometric', 'reliability', 'validity') and instrument types (eg, 'questionnaire', 'scale', 'self-report'). An experienced information specialist was consulted to refine these strategies, which were subsequently tested to verify their ability to capture key papers already identified in the field. The strategies have been carefully designed to balance sensitivity and specificity, ensuring comprehensive identification of relevant studies while maintaining feasibility. The complete search strategies have been uploaded to PROSPERO as part of the registration process and are available in the online supplemental materials.

Study records

Data management

Literature search results will be imported into EndNote (V.21) reference management software, which will be used to organise and manage citations throughout the review process. We will identify and remove duplicate references using EndNote's built-in duplicate detection feature, followed by manual screening by the investigators. After deduplication, the Rayyan platform[30] will be used to screen titles and abstracts.

For data extraction and synthesis, we will use the standardised 'COSMIN Review Management File', a Microsoft Excel template specifically designed for systematic reviews of measurement properties and provided by the COSMIN initiative. The management file facilitates structured data extraction, organisation of measurement property evidence, methodological quality assessment and synthesis of findings across studies for each instrument. However, this template may be modified as necessary during the review process to accommodate emerging data patterns, with any changes documented in the final report.

Table 1 Eligibility criteria

Criteria	Inclusion criteria	Exclusion criteria
Population	Studies involving human adults (18 years or older) from either clinical or non-clinical populations.	Studies with participants younger than 18 years. Studies focusing exclusively on highly specific adult subgroups (eg, parents, specific occupational groups) where the measure is validated only within that narrow context.
Instruments	Self-report measures (eg, questionnaires, scales) specifically designed to assess mentalising, mentalisation or reflective functioning in adults as a general capacity. Modified versions of established measures are included if the modification is described and the core construct remains mentalising.	Measures of neighbouring but distinct constructs (eg, empathy, perspective-taking, psychological mindedness). Measures designed to assess mentalising only within a highly specific, narrow context (eg, mentalising about a single relationship type).
Comparators	Not applicable. This review assesses the measurement properties of instruments, not the comparison of interventions.	Not applicable.
Outcomes	Studies must report on at least one of the following measurement properties: content validity, structural validity, internal consistency, cross-cultural validity/measurement invariance, reliability, measurement error, criterion validity, hypotheses testing for construct validity, or responsiveness.	Studies that do not report any psychometric data for the mentalising measure. Studies that use a mentalising measure solely to validate a different instrument without reporting on the properties of the mentalising measure itself.
Study design	Any study design that reports on the development or evaluation of measurement properties. Eligible publication types include peer-reviewed journal articles, dissertations, theses and unpublished materials.	Review papers, case studies, conference abstracts and theoretical papers without empirical data.
Language	English-language publications (due to resource constraints for translation and to ensure consistent quality assessment by the review team).	Publications in languages other than English.
Publication date	No restriction by publication date (from inception onwards).	Not applicable.

To maintain methodological rigour and ensure standardised evaluation across instruments, our analysis will be based exclusively on data reported in published documents. This approach aligns with the COSMIN methodology and ensures reproducibility by relying on publicly accessible information. Additionally, this standardised approach prevents potential variability in evidence synthesis that could arise from inconsistent author responses to queries about missing information.

Selection process

Titles and abstracts of all retrieved references will be screened against the eligibility criteria. Full texts of potentially eligible studies will then be retrieved and assessed to determine final inclusion. We will document the selection process in a PRISMA flow diagram,[31] recording the number of studies identified, screened, assessed for eligibility and included in the review, along with reasons for exclusions at the full-text screening stage.

Review process and disagreement resolution

Throughout this systematic review, two reviewers (AA and AHDK) will independently perform all key methodological steps (including screening by titles and abstracts, full-text assessment, data extraction and quality assessment). Before beginning the full screening process, a pilot test of five papers will be conducted to ensure reviewers are familiar with the eligibility criteria and to calibrate their assessment approach. When disagreements arise at any stage, reviewers will first discuss discrepancies to reach a consensus. If consensus cannot be achieved, a third reviewer with expertise in mentalising theory and psychometrics will be consulted to make a final determination.

If amendments to this protocol become necessary during the course of the review, the date of each amendment will be recorded along with a description of the change and the rationale for it.

Data items

The following data will be extracted from each eligible study:

1. Measure characteristics: name and version of the instrument; construct(s) being measured; original development article; target population for which the measure was developed; context of use; number of scales/subscales and items; response options; original language and available translations and whether the measure is based on a reflective or formative model.
2. Study population characteristics: sample size; age (mean, SD, range); gender distribution; country/language; setting (clinical/non-clinical) and inclusion/exclusion criteria.

3. Results on measurement properties: content validity (eg, relevance, comprehensiveness and comprehensibility); structural validity (eg, factor structure, item loadings, fit indices); internal consistency (eg, alpha, omega); cross-cultural validity/measurement invariance (if applicable); reliability (eg, intraclass coefficient, kappa, test–retest correlations); measurement error (if reported); criterion validity (considering the RFS as the gold standard); hypotheses testing for construct validity (eg, correlations with other instruments, known-groups comparisons); responsiveness; feasibility aspects (availability and cost) and interpretability aspects (distribution of scores and normative values).

Risk of bias in individual studies

The methodological quality of included studies will be assessed using the COSMIN Risk of Bias checklist.[28] Table 2 presents the summarised assessment criteria for each measurement property. Each property will be rated on a four-point scale (very good, adequate, doubtful or inadequate), with the 'worst score counts' principle applied whereby the lowest rating of any criterion determines the overall quality rating for that property.

Data synthesis

Data synthesis will follow the COSMIN methodology for systematic reviews of PROMs: (1) For each measurement property of each measure, we will rate the results of each study against the criteria for good measurement properties as sufficient (+), insufficient (−) or indeterminate (?); (2) We will summarise the evidence across studies and rate the summarised result, classifying them as (+), insufficient (−), inconsistent (±) or indeterminate (?); (3) Using a modified Grading of Recommendations Assessment, Development and Evaluation (GRADE)[32] approach, considering risk of bias, inconsistency, imprecision and indirectness, we will grade the quality of evidence as high, moderate, low or very low; and (4) We will formulate recommendations for the use of mentalising self-report measures in adults using the following categories:

- *Category A.* Measures with sufficient content validity and at least low-quality evidence for sufficient internal consistency and other measurement properties,
- *Category B.* Measures with sufficient content validity but limited evidence on other measurement properties,
- *Category C.* Measures with high-quality evidence for an insufficient measurement property,
- *Category D.* Measures with indeterminate ratings on measurement properties due to limited evidence.

When inconsistent results are found across studies for the same measurement property of a measure, we will follow the COSMIN approach to inconsistency management.[28] First, we will seek explanations for inconsistency (eg, differences in study populations, methodological quality, or measurement conditions). If explanations are identified, we will summarise results separately for relevant subgroups or focus only on high-quality studies. If no explanations are found, we will either rate the measurement property as inconsistent (±) or base our rating on the majority of consistent results while downgrading the quality of evidence for inconsistency in our GRADE assessment.

In addition to evaluating published content validity studies, our review team will independently assess each mentalising measure's content validity by rating its relevance, comprehensiveness and comprehensibility according to the COSMIN criteria for good content validity. This assessment provides an additional source of evidence that is especially valuable when content validity studies are limited or non-existent. When summarising the overall evidence on content validity, we will consider three sources: the PROM development process, available content validity studies and our review team's structured assessment.[28]

We will provide a narrative synthesis of the findings. Results will be presented in summary tables, including a comprehensive summary of findings table displaying the evidence synthesis for each measurement property of each measure.

Patient and public involvement

No patients or members of the public were involved in the design or conduct of this protocol. This was due to the highly technical and methodological focus of the work, which is centred on the application of psychometric standards (the COSMIN methodology) to evaluate measurement instruments.

Ethics and dissemination

Formal ethical approval is not required for this systematic review, as it will be based on the analysis of previously published and publicly available data. No primary data will be collected, and no human participants will be directly involved. The findings of this systematic review will be disseminated broadly to ensure they reach relevant academic, clinical and public audiences.

DISCUSSION

This systematic review will provide a comprehensive evaluation of self-report measures of mentalising for adults, creating an inventory of available instruments and assessing their alignment with contemporary conceptualisations of the construct.

The application of the COSMIN methodology provides a robust framework for this evaluation. However, beyond the specific methodological limitations noted previously, a broader challenge involves the evolving nature of the mentalising construct itself over the past two decades.[33 34] To address this, we will carefully map each instrument's content against current theoretical frameworks to provide a transparent assessment of conceptual coverage.

Additional limitations of the evidence base may include: (1) the quality of primary studies, which may

Table 2 Overview of the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) risk of bias assessment criteria by measurement property

Measurement property	Objective	Key assessment criteria
PROM development	To evaluate the rigour of the initial instrument development process.	► *Concept elicitation*: Assesses if relevant concepts were identified from a sample representing the target population using appropriate qualitative methods (eg, interviews) and if data collection continued until saturation was reached. ► *Pilot testing*: Checks if the comprehensibility of instructions, items and response options was tested in the target population using a suitable method (eg, cognitive interviews).
Content validity	To assess if the PROM's content is an adequate reflection of the construct to be measured.	► *Relevance and comprehensiveness*: Evaluates whether patients and/or professionals were asked if the items are relevant and if they cover all important aspects of the construct. ► *Comprehensibility*: Assesses whether the clarity of items and instructions was confirmed with patients and/or professionals. ► *Methodology*: Requires appropriate qualitative or quantitative methods with adequate sample sizes.
Structural validity	To determine if the scores adequately reflect the dimensionality of the construct.	► *Statistical method*: Prefers Confirmatory Factor Analysis ('very good') over exploratory factor analysis ('adequate'). Considers principal component analysis (PCA) 'doubtful'. For item response theory, it checks if the model fits the research question. ► *Sample size*: Must be adequate for the analysis (eg, for factor analysis, ≥7 participants per item and ≥100 total is 'very good').
Internal consistency	To measure the degree of interrelatedness among the items in a (sub)scale.	► *Statistical calculation*: Verifies that an appropriate statistic was calculated for the score type, such as Cronbach's alpha or omega for continuous scores, Kuder-Richardson 20 (KR-20) for dichotomous scores, or the SE of theta (SE(θ)) for IRT-based scores.
Cross-cultural validity/measurement invariance	To assess whether PROM items perform equivalently across different groups (eg, cultures, languages).	► *Sample comparability*: Checks if the groups were similar on relevant characteristics, except for the grouping variable itself. ► *Statistical method*: Requires an appropriate approach, such as multi-group confirmatory factor analysis (MGCFA) or IRT-based differential item functioning (DIF) analysis. ► *Sample size*: Must be adequate for the analysis (eg, for MGCFA, ≥7 subjects per item and ≥100 per group).
Reliability	To determine the extent to which scores are consistent for stable patients over repeated measurements.	► *Design requirements*: The study design must ensure that (1) patients were stable in the construct being measured, (2) the time interval was appropriate and (3) measurement conditions were similar. ► *Statistical method*: Requires an appropriate statistic, such as an intraclass correlation coefficient (ICC) based on an agreement model for continuous scores or a weighted kappa for ordinal scores.
Measurement error	To quantify the systematic and random error in a score that is not attributable to true change.	► *Design requirements*: Same as for reliability (stable patients, appropriate interval, similar conditions). ► *Statistical method*: Assesses whether the SE of measurement (SEM), smallest detectable change (SDC) or limits of agreement (LoA) were calculated, preferably using an agreement model.
Criterion validity	To evaluate the degree to which PROM scores are an adequate reflection of a 'gold standard'.	► *Statistical method*: Checks for the use of appropriate statistics to compare the PROM to the gold standard, such as correlations, area under the curve (AUC) or sensitivity and specificity.
Hypotheses testing for construct validity	To assess if PROM scores behave in accordance with predefined hypotheses about the construct.	► *Comparator instruments*: For convergent validity, the construct and measurement properties of the comparator instrument(s) must be clear and adequate. ► *Known groups*: For discriminative validity, the subgroups being compared must be clearly described. ► *Statistical method*: Must be appropriate for the comparison being made.

Continued

Table 2 Continued

Measurement property	Objective	Key assessment criteria
Responsiveness	To measure the PROM's ability to detect change over time in the construct.	► *Criterion approach*: Compares change scores on the PROM to change scores on a gold standard using appropriate statistics (eg, correlations, AUC). ► *Construct approach*: Tests hypotheses about expected changes, such as comparing change scores with other instruments or evaluating changes in subgroups before and after an intervention.

IRT, Item response theory; PROM, Patient-reported outcome measure.

not have been designed to meet COSMIN standards; (2) potential language bias from excluding non-English publications, which may limit the generalisability of findings particularly for measures developed in non-English speaking countries; (3) the possibility of selective outcome reporting within included studies and (4) the challenge of comparing measures developed under different theoretical frameworks of mentalising. These limitations will be explicitly considered when formulating our recommendations. For instance, if our review reveals that evidence for most instruments is of low or very low quality due to widespread methodological flaws in primary studies, this will temper recommendations for their clinical use and highlight an urgent need for more rigorous psychometric research. Thus, it remains uncertain whether the evidence will permit a definitive recommendation of a single 'best' instrument, but the synthesis will provide a clear evidence base to guide future instrument selection and development.

Contributors AA conceived of the study, developed the search strategy, and drafted the manuscript. The protocol methodology was designed by AA and AHDK. MAM and SG supervised the project. All authors approved the final version of the protocol. AA serves as the guarantor for this work.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Ahmad Asgarizadeh http://orcid.org/0000-0001-8431-0548

REFERENCES

1 Luyten P, Fonagy P. The neurobiology of mentalizing. *Personal Disord* 2015;6:366–79.
2 Bateman A, Fonagy P. *Mentalization-based treatment for personality disorders: a practical guide*. New York, NY, US: Oxford University Press, 2016.
3 Johnson BN, Kivity Y, Rosenstein LK, *et al*. The association between mentalizing and psychopathology: a meta-analysis of the reading the mind in the eyes task across psychiatric disorders. *Clin Psychol: Sci Pract* 2022;29:423–39.
4 Katznelson H. Reflective functioning: a review. *Clin Psychol Rev* 2014;34:107–17.
5 McLaren V, Gallagher M, Hopwood CJ, *et al*. Hypermentalizing and borderline personality disorder: a meta-analytic review. *Am J Psychother* 2022;75:21–31.
6 Bora E. A meta-analysis of theory of mind and "mentalization" in borderline personality disorder: a true neuro-social-cognitive or meta-social-cognitive impairment? *Psychol Med* 2021;51:2541–51.
7 Sloover M, van Est LAC, Janssen PGJ, *et al*. A meta-analysis of mentalizing in anxiety disorders, obsessive-compulsive and related disorders, and trauma and stressor related disorders. *J Anxiety Disord* 2022;92:102641.
8 Simonsen CB, Jakobsen AG, Grøntved S, *et al*. The mentalization profile in patients with eating disorders: a systematic review and meta-analysis. *Nord J Psychiatry* 2020;74:311–22.
9 Chung YS, Barch D, Strube M. A meta-analysis of mentalizing impairments in adults with schizophrenia and autism spectrum disorder. *Schizophr Bull* 2014;40:602–16.
10 Fonagy P, Target M, Steele H, *et al*. *Reflective-functioning manual, Version 5.0, for application to adult attachment interviews*. London: University College London, 1998.
11 Taubner S, Hörz S, Fischer-Kern M, *et al*. Internal structure of the reflective functioning scale. *Psychol Assess* 2013;25:127–35.
12 Achim AM, Guitton M, Jackson PL, *et al*. On what ground do we mentalize? Characteristics of current tasks and sources of information that contribute to mentalizing judgments. *Psychol Assess* 2013;25:117–26.
13 Pinkham AE, Penn DL, Green MF, *et al*. Social cognition psychometric evaluation: results of the initial psychometric study. *Schizophr Bull* 2016;42:494–504.
14 Wendt LP, Zimmermann J, Spitzer C, *et al*. Mindreading measures misread? A multimethod investigation into the validity of self-report and task-based approaches. *Psychol Assess* 2024;36:365–78.
15 Fonagy P, Luyten P, Moulton-Perkins A, *et al*. Development and validation of a self-report measure of mentalizing: the reflective functioning questionnaire. *PLoS One* 2016;11:e0158678.
16 Hausberg MC, Schulz H, Piegler T, *et al*. Is a self-rated instrument appropriate to assess mentalization in patients with mental disorders? Development and first validation of the mentalization questionnaire (MZQ). *Psychother Res* 2012;22:699–709.
17 Dimitrijević A, Hanak N, Altaras Dimitrijević A, *et al*. The mentalization scale (MentS): a self-report measure for the assessment of mentalizing capacity. *J Pers Assess* 2018;100:268–80.
18 Müller S, Wendt LP, Spitzer C, *et al*. A critical evaluation of the reflective functioning questionnaire (RFQ). *J Pers Assess* 2022;104:613–27.

19 Asgarizadeh A, Ghanbari S. Iranian adaptation of the Epistemic Trust, Mistrust, and Credulity Questionnaire (ETMCQ): validity, reliability, discriminant ability, and sex invariance. *Brain Behav* 2024;14:e3455.
20 Asgarizadeh A, Sharp C, Ghanbari S. Shame-coping clusters: comparisons regarding attachment insecurities, mentalizing deficits, and personality pathology, controlling for general emotion dysregulation. *Borderline Personal Disord Emot Dysregul* 2023;10:25.
21 Blay M, Bouteloup M, Duarte M, *et al*. Association between pathological narcissism and emotion regulation: the role of self-mentalizing? *Personal Ment Health* 2024;18:227–37.
22 Duschinsky R, Foster S. *Mentalising and epistemic trust: The work of Peter Fonagy and colleagues at the Anna Freud Centre*. New York: Oxford University Press, 2021.
23 Fonagy P, Luyten P. A developmental, mentalization-based approach to the understanding and treatment of borderline personality disorder. *Dev Psychopathol* 2009;21:1355–81.
24 Humfress H, O'Connor TG, Slaughter J, *et al*. General and relationship-specific models of social cognition: explaining the overlap and discrepancies. *J Child Psychol Psychiatry* 2002;43:873–83.
25 Vizgaitis AL, Lenzenweger MF. Identity pathology and mentalization deficits: An attempt to support clinical theory with data. *Pers Disord Theory Res Treat* 2024;15:128–33.
26 Kivity Y, Levy KN, Johnson BN, *et al*. Mentalizing in and out of awareness: a meta-analytic review of implicit and explicit mentalizing. *Clin Psychol Rev* 2024;108:102395.
27 Prinsen CAC, Mokkink LB, Bouter LM, *et al*. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018;27:1147–57.
28 Mokkink LB, Elsman EBM, Terwee CB. COSMIN guideline for systematic reviews of patient-reported outcome measures version 2.0. *Qual Life Res* 2024;33:2929–39.
29 Moher D, Shamseer L, Clarke M, *et al*. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev* 2015;4:1.
30 Ouzzani M, Hammady H, Fedorowicz Z, *et al*. Rayyan-a web and mobile app for systematic reviews. *Syst Rev* 2016;5:210.
31 Moher D, Liberati A, Tetzlaff J, *et al*. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med* 2009;6:e1000097.
32 Guyatt G, Oxman AD, Akl EA, *et al*. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol* 2011;64:383–94.
33 Fonagy P, Luyten P, Allison E, *et al*. What we have changed our minds about: Part 2. Borderline personality disorder, epistemic trust and the developmental significance of social communication. *Borderline Personal Disord Emot Dysregul* 2017;4:9.
34 Fonagy P, Luyten P, Allison E, *et al*. What we have changed our minds about: Part 1. Borderline personality disorder as a limitation of resilience. *Borderline Personal Disord Emot Dysregul* 2017;4:11.