

Review

Construct validity evidence reporting practices for the Reading the Mind in the Eyes Test: A systematic scoping review

Wendy C. Higgins^{a,*}, David M. Kaplan^a, Eliane Deschrijver^{a,b}, Robert M. Ross^c

^a Macquarie University, School of Psychological Sciences, NSW 2109, Australia

^b University of Sydney, School of Psychology, NSW 2006, Australia

^c Macquarie University, Department of Philosophy, NSW 2109, Australia

ARTICLE INFO

Keywords:

Validity
Reading the Mind in the Eyes Test
Measurement
Construct validation
Social cognition

ABSTRACT

The Reading the Mind in the Eyes Test (RMET) is one of the most influential measures of social cognitive ability, and it has been used extensively in clinical populations. However, questions have been raised about the validity of RMET scores. We conducted a systematic scoping review of the validity evidence reported in studies that administered the RMET ($n = 1461$; of which 804 included at least one clinical sample) with a focus on six key dimensions: internal consistency, test-retest reliability, factor structure, convergent validity, discriminant validity, and known group validity. Strikingly, 63% of these studies failed to provide validity evidence from any of these six categories. Moreover, when evidence was reported, it frequently failed to meet widely accepted validity standards. Overall, our results suggest a troubling conclusion: the validity of RMET scores (and the research findings based on them) are largely unsubstantiated and uninterpretable. More broadly, this project demonstrates how unaddressed measurement issues can undermine a voluminous psychological literature.

The Reading the Mind in the Eyes Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) is one of the most widely used measures of social cognitive ability in both clinical and non-clinical populations (Eddy, 2019; Pavlova & Sokolov, 2022). The test comprises 36-items, each of which presents a black and white photo of a White person's eyes with four mental state descriptors (Fig. 1). The task is to select the descriptor that best matches what the person in the photo is thinking or feeling.

The journal article that introduced the RMET has been cited at least 7412 times (Google Scholar as of November 19, 2023), and the perceived utility of the RMET as a measure of social cognitive ability is evidenced in a variety of ways. First, it is included in The United States of America's National Institute of Mental Health's (NIMH) Research Domain Criteria Initiative as a "current best option" for a task to assess the ability to understand mental states (NIMH, 2016). Second, it has been used in high stakes clinical contexts, including brain mapping during awake brain surgery as a tool to avoid resection of brain areas involved in social cognition (Herbet, Lafargue, Moritz-Gasser, Bonnet-blanc, & Duffau, 2015; Herbet, Rigaux-Viodé, & Moritz-Gasser, 2017; Herbet & Moritz-Gasser, 2019). Third, it has been used in trials of

pharmacological interventions to improve social cognitive ability (Zamoscic et al., 2021; Zhong et al., 2021). Fourth, in a survey of experts in social cognition, the RMET was nominated as one of the best existing measures of social cognition for use in schizophrenia research (Pinkham et al., 2014).¹ Fifth, it is one of the most frequently administered tests of social cognitive ability in neuropsychiatric research (Eddy, 2019). Moreover, its widespread use is likely facilitated by its being freely available, easy to administer (both as a pen and paper test and as an online test), and easy to score.

Unfortunately, serious concerns have been raised about the construct validity of RMET scores (Higgins, Ross, Polito, & Kaplan, 2023; Kittel, Olderbak, & Wilhelm, 2022; Silverman, 2022). Construct validity refers to the extent to which test scores can be interpreted as measuring what they are intended to measure (Clark & Watson, 2019; Furr, 2022). Despite the common practice of referring to tests as "valid", validity is not a property of a test, but rather an assessment of the interpretability of a test score for a particular use in a particular sample (American Educational Research Association [AERA] et al., 2014; Appelbaum et al., 2018; Clark & Watson, 2019; Furr, 2022). Because the psychometric properties of test responses can vary significantly between groups

* Corresponding author at: Macquarie University, NSW 2109, Australia.

E-mail address: wendy.higgins@mq.edu.au (W.C. Higgins).

¹ However, it is worth noting that in subsequent research involving participants diagnosed with schizophrenia Pinkham and colleagues found that other social cognition tasks have better psychometric properties than the RMET (Pinkham, Harvey, & Penn, 2018; Pinkham et al., 2016).

(especially between clinical versus non-clinical groups), establishing construct validity should be an ongoing process in which test users routinely justify their choice of test by referencing existing validity evidence specific to their proposed test use and target populations and provide new validity evidence demonstrating that the test scores are valid within their samples (AERA et al., 2014; Appelbaum et al., 2018; Clark & Watson, 2019). Critically, the interpretability of psychological research findings depends on the extent to which construct validity evidence supports the measurements used within a given study (Flake & Fried, 2020; Flake, Pek, & Hehman, 2017; Furr, 2022; Vazire, Schiavone, & Bottesini, 2022), rendering weak or absent validity evidence highly problematic.

Construct validity concerns for the RMET include substantial limitations of the methods used to construct and initially validate the test (described below), uncertainty over which aspect of social cognition the test measures (e.g., emotion recognition, theory of mind, or empathy; Kittel et al., 2022; Oakley, Brewer, Bird, & Catmur, 2016), uncertain factor structure (Higgins, Ross, Langdon, & Polito, 2023; Olderbak et al., 2015; Van Staden & Callaghan, 2021), low levels of internal consistency (De Paoli, Fuller-Tyszkiewicz, Huang, & Krug, 2020; Vellante et al., 2013), failures to find statistically significant correlations with other measures of social cognitive ability (Gernsbacher & Yergeau, 2019), and evidence that RMET scores may be influenced by social class (Dodell-Feder, Ressler, & Germine, 2020), social norms (Johnston, Miles, & McKinlay, 2008), culture (Dodell-Feder et al., 2020; Van Staden & Callaghan, 2021), vocabulary (Gernsbacher & Yergeau, 2019; Kittel et al., 2022), education (Warrier et al., 2018), and IQ scores (Baker, Peterson, Pulos, & Kirkland, 2014; Peñuelas-Calvo, Sareen, Sevilla-Llewellyn-Jones, & Fernández-Berrocal, 2019).

The validity evidence reported in the RMET's initial validation study (Baron-Cohen, Wheelwright, Hill, et al., 2001) has significant limitations (Johnston et al., 2008; Silverman, 2022). Of particular concern, the structural properties of RMET scores, including their factor structure and internal consistency, were not reported (Higgins, Ross, Polito, & Kaplan, 2023). Rather, key sources of validity evidence that were reported included (1) lower average RMET scores for 15 male autistic individuals² compared to controls and (2) statistically significant

negative correlations between RMET scores and scores on a self-report measure of autistic traits. However, these sources of validity evidence both rely heavily on Baron-Cohen's (1995) theory that autistic people have a specific deficit in the social cognitive capacity of theory of mind (ToM), which has been contested (Deschrijver & Palmer, 2020; Gernsbacher & Yergeau, 2019; Milton, Gurbuz, & López, 2022; Silverman, 2022).

The method by which the target response for each RMET item was established also raises questions about the construct validity of RMET scores. As described in the initial validation study (Baron-Cohen, Wheelwright, Hill, et al., 2001), the photos in the RMET were taken from magazines and the "correct" and "incorrect" responses for each RMET item were investigator-determined. RMET items were then verified in a sample comprising undergraduate students from the University of Cambridge ($n = 103$) and members of the general public in Cambridge and Exeter ($n = 122$), with the consensus criteria being that at least 50% of participants selected the "correct" response and no more than 25% of them selected the same "incorrect" response. This method of verifying the "correct" responses means that the generalisability of the target responses relies on the generalisability of the responses in this sample, approximately half of which consisted of students from an elite university. Item level responses for the RMET's four mental state descriptors have rarely been reported in subsequent research (Higgins, Ross, Polito, & Kaplan, 2023), but when they are reported there are frequently multiple items that fail to meet the initial consensus criteria (e.g., Eddy & Hansen, 2020; Higgins, Ross, Langdon, & Polito, 2023; Olderbak et al., 2015; Prevost et al., 2014; Van Staden & Callaghan, 2021). These findings suggest that the "correct" responses might not routinely generalise across samples, limiting the interpretability of RMET scores as a measure of social cognitive ability in diverse populations without additional validity evidence. We describe limitations of the initial validation study in detail in section S1 of the supplementary materials.

Several researchers have raised concerns about the validity of RMET scores over the past 15 years (for early critiques see Johnston et al., 2008; Peterson & Miller, 2012). Nevertheless, use of the RMET is widespread and appears to be increasing (Pavlova & Sokolov, 2022),

a) Item 17

doubtful

affectionate



playful

aghast

b) Item 34

aghast

baffled



distrustful

terrified

Fig. 1. Sample RMET Items. *Note.* The RMET is freely available for download from University of Cambridge's Autism Research Centre's website at <https://www.autismresearchcentre.com/tests/eyes-test-adult/>. The target response for item 17 is "doubtful", and the target response for Item 34 is "distrustful".

² We use identity first terminology to refer to autistic individuals. While there are mixed language preferences within the autism community, it has been argued that identity first language tends to be more preferred than person first language (Botha, Hanlon, & Williams, 2023).

suggesting that the prevailing attitude is that these concerns are not serious and that the RMET is a valuable scientific tool. The apparent disregard of concerns about the validity of RMET scores is consistent with a growing body of evidence demonstrating that researchers in psychological science routinely fail to pay serious attention to construct validity evidence (Flake et al., 2017; Flake, Davidson, Wong, & Pek,

2022; Flake & Fried, 2020; Fried, Flake, & Robinaugh, 2022; Hussey & Hughes, 2020; Schimmack, 2021; Shaw, Cloos, Luong, Elbaz, & Flake, 2020; Slaney, 2017; Vassar & Hale, 2009). This is of concern because rigorous psychological science requires the reporting of construct validity evidence, and it is therefore critical that the validity evidence for scores on measures as widely used as the RMET are comprehensively surveyed. To achieve this, we conduct a scoping review. Scoping reviews provide systematic, transparent, reproducible, and reliable information syntheses that can be used to address broad research questions, “identify the types of available evidence in a given field” and “identify and analyze gaps in the knowledge base” (Munn et al., 2018, pp. 3–4).

In the present review we focus on a key overarching question: To what extent do empirical studies³ in the RMET literature provide validity evidence to justify the administration of the RMET and to substantiate the interpretation of RMET scores? In particular, we examine the frequency with which these studies have reported six key dimensions of construct validity evidence: test-retest reliability, internal consistency, factor structure, convergent validity, discriminant validity, and known group validity. Additionally, we examine three variations in how the RMET has been used in practice that have implications for validity, including (1) modifications to the RMET (e.g., translations and abbreviations), (2) populations that are studied using the RMET (e.g., different clinical populations), and (3) latent psychological constructs that the RMET is used to measure (e.g., cognitive empathy and ToM). Finally, we synthesise this evidence to evaluate the extent to which it supports interpretations of RMET scores as measurements of social cognitive ability.

1. Method

1.1. Transparency and openness

We adhered to the PRISMA 2020 guidelines for systematic reviews (Page et al., 2021) using the extension for scoping reviews (Tricco et al., 2018). This review was preregistered on the Open Science Framework (OSF; accessible at <https://osf.io/9bujh/registrations>). All data have been made publicly available on the project's OSF page (accessible at <https://osf.io/9bujh/>).

1.2. Inclusion/exclusion criteria

We included studies that met the following criteria: (1a) reported an original empirical study that analysed RMET scores or (1b) were meta-analyses, systematic reviews, or scoping reviews that included at least one study that administered the RMET; (2) were published between 2001 and 2021 inclusive⁴; (3) the full text was available in English; and (4) the full text was accessible to us (i.e., accessible online or through our university library's catalogue for direct access or interlibrary loan). The key focus of our project is multi-subject empirical studies, which we summarise and discuss in the main text. However, validity evidence is also important for meta-analyses and reviews that report findings based on RMET scores (Higgins, Ross, Polito, & Kaplan, 2023), and we summarise validity evidence reporting practices for the other publication types in supplementary materials (see section S2 and Table S1).

We limited our review to studies that used the revised version of the RMET introduced in Baron-Cohen, Wheelwright, Hill, et al. (2001). We

excluded studies that used the original version introduced in Baron-Cohen, Jolliffe, Mortimore, and Robertson (1997) because the revised version was designed by Baron-Cohen and colleagues to address substantial weaknesses in the original test, including ceiling effects (see supplementary materials section S1 for details). We also excluded the child version of the RMET introduced in Baron-Cohen, Wheelwright, Spong, Scahill, and Lawson (2001) because it had been adapted for use with children.

1.3. Search strategy

We searched the Web of Science and Scopus databases using abstract, title, and keywords search with the query ““Mind in the Eye*” OR “eye* test” OR “eye* task”” from 2001 to 2021. Because the measures used in a study are often not mentioned in the abstract, title, or keywords, we also conducted a full text search using the same search query in EBSCO's Academic Search Premier research database (<https://www.ebsco.com/products/research-databases/academic-search-premier>). Studies that were listed in an included systematic review as administering the RMET or cited in any included study as a source of validity evidence for RMET scores were also screened for inclusion. After screening the initial search results, we observed that our search strategy had missed studies that we knew included the RMET, which we concluded was due to the limited number of publications available for full text search in the Academic Search Premier database. Consequently, to increase the comprehensiveness of our search, we deviated from our pre-registration and added an additional search strategy: using the “cited by” option in Web of Science and Scopus, we identified studies that cited the initial validation study of the RMET (i.e., Baron-Cohen, Wheelwright, Hill, et al., 2001). The final database search was conducted on March 29, 2022.⁵ Studies we identified were imported to *Covidence systematic review software* (Veritas Health Innovation, 2022), a web-based collaboration software platform that streamlines the production of systematic literature reviews, which we used for screening and data extraction.

1.4. Screening and data extraction

Title, abstract, and full text screening were conducted by one investigator (WCH). Records were excluded if they did not meet the criterion for type of publication or if the content in the abstract made it clear that the “eye test” referred to in the publication was not the RMET (e.g., it was a measure of visual acuity, eye movement, or eyewitness recall). In cases of uncertainty, records were progressed to full text screening and assessed on the inclusion and exclusion criteria outlined above.

A data extraction template was created in *Covidence*. The full template is available on the study's OSF page (<https://osf.io/9bujh/>). The key variables extracted were: (1) bibliographic information (author, publication year, title), (2) language version(s) of the RMET, (3) construct(s) that the RMET was purportedly measuring, (4) population sampled and sample size,⁶ (5) modifications to the test (e.g., abbreviated versions) or test scoring (e.g., subscores) and (6) validity evidence (internal consistency, test-retest reliability, factor analysis, convergent validity, discriminant validity, and known group validity). For each type of validity evidence, we coded whether (i) a statistic was calculated using new participant data, (ii) evidence was cited from previous

³ For the purposes of our review, a “study” refers to a publication that reports findings from one or more empirical studies that use the RMET.

⁴ Studies that were only available as “online first” publications at the time of screening were included in the scoping review and were treated as being published in 2021 (including in Figure 2). Some of them may have been subsequently assigned to journal issue with a 2022 publication date. Dissertations, theses and pre-prints that had not been accepted for publication were not included in the scoping review.

⁵ To ensure that our review remains up to date, Scopus and Web of Science searches for studies citing Baron-Cohen, Wheelwright, Hill, et al. (2001) published from 2001 to 2021 were also conducted September 20, 2023. The 10 most recently added studies in each database were checked, and all studies had already been identified in our review with the exception of one, which we exclude because the full text is in German.

⁶ Sample sizes can be found in the “Full dataset” spreadsheet on the study's OSF page.

research, (iii) evidence was calculated using new participant data and cited from previous research, or (iv) no information was provided. We coded studies as having “cited” evidence even if they did not include an explicit citation to previous research. For example, if a study stated that the RMET has good test-retest reliability, the study was coded as citing validity evidence from previous research even if a citation to support the statement was not provided. Where provided, the statistics for test-retest reliability, internal consistency, and factor analyses were also recorded in *Covidence*. Section S3 of the supplementary materials includes a description of each type of extracted validity evidence and summaries of common methods for evaluating them; we include widely used cut-off values for statistics where applicable as indicators of values that have been conventionally accepted to support validity.⁷

Data from all studies were extracted by WCH.⁸ In addition, data were independently extracted and coded by a research assistant for a random subset (15% [$n = 218$]) of studies. Inter-rater reliability was assessed via percentage agreement and Bennett, Alpert, and Goldstein's (1954) s -statistic, which can provide a more useful estimate of inter-rater reliability for categorical variables as it takes chance agreement into account (Feinstein & Cicchetti, 1990).⁹ Twelve studies that were used for training the research assistant and to calibrate data extraction were excluded from the inter-rater reliability calculations. Conflicts were resolved by rechecking the full text document and subsequent discussion between the coders when required.

Our primary interest is in the extent to which academics are using construct validity evidence to substantiate their interpretations of RMET scores. Consequently, we coded convergent and discriminant validity evidence if, anywhere in a study, there was explicit discussion of the association between scores on two tests in the context of evaluating what the RMET measures (even if assessing convergent and discriminant validity were not explicitly mentioned). Similarly, we coded known group validity evidence if, anywhere in a study, there was an explicit discussion of the performance of a group relative to another group in the context of assessing what the RMET measures (even if assessing known group validity was not explicitly mentioned). Whenever the differentiation of groups or correlations with other relevant measures were included in the description of the RMET in the methods section, we also coded these types of validity evidence even if an explicit link was not drawn to what the RMET measures, under the assumption that this information was provided in support of the inclusion of the RMET in the study.

2. Results

2.1. Inter-rater reliability

For all categories of validity evidence that we examined, inter-rater reliability was high according to both percentage agreement and the s -statistic (see Table S3). Levels of agreement were lowest for the coding

of “modified”. In this instance, we provided a dichotomous response to indicate whether or not a study used a modified version of the RMET or a modified scoring method. However, this response captures numerous possible types of modifications. For example, scoring modifications included the use of subscores, z -scores, or composite scores with other measures of social cognition, while modifications to the test itself included reducing the number of test items and reducing the number of response choices. While the s -statistic is low (0.61), the percentage agreement (0.81) indicates that, despite the multitude of potential modifications, there was a high level of agreement about whether or not a study included any modifications. Notably, the inter-rater reliability for specific modifications (i.e., subscores, fewer items, fewer response choices) was high.

2.2. Included studies

We found an upward trend in the number of studies produced each year that used the RMET (Fig. 2), identifying a total of 1461 studies that met our inclusion criteria (Fig. 3).

2.3. Construct validity evidence

2.3.1. Test-retest reliability

The test-retest reliability of RMET scores was calculated in 1.4% ($n = 21$) of studies and 8.4% ($n = 122$) of studies cited evidence from previous research, while 90.8% ($n = 1327$) of studies did not provide any evidence.¹⁰ Calculated test-retest reliability was generally acceptable, with point estimates of correlations above 0.60 in 20 out of 21 studies (see Table S4). Consistent with this, only one study that calculated test-retest reliability concluded that RMET scores have poor test-retest reliability (Chen et al., 2017), while 77.3% ($n = 17$) reported acceptable to good test-retest reliability, and 13.6% ($n = 3$) did not provide an interpretation of the analysis. In 56.6% ($n = 69$) of studies that cited evidence from previous research, the language version administered differed for the cited versus citing studies. This may be due to the unavailability of test-retest reliability evidence specific to the target language version. In particular, while we identified eight studies that have evaluated the test-retest reliability of scores on the English language RMET; remarkably, the earliest of these studies to assess test-retest reliability in a non-clinical sample was not published until 2016 (Pinkham, Penn, Green, & Harvey, 2016), a full 15 years after the RMET was first published.

We did not systematically code whether or not studies cited as evidence for the test-retest reliability of RMET scores used the same language version of the test as the citing studies; nonetheless, we did observe that the most frequently cited studies administered Italian (Vellante et al., 2013, cited in $n = 61$ studies, only 14 of which used an Italian RMET), Spanish (Fernandez-Abascal, Cabello, Fernandez-Berrocá, & Baron-Cohen, 2013, cited in $n = 36$, only seven of which used a Spanish RMET), and Swedish language versions of the RMET (Hallerbäck, Lugnegård, Hjärthag, & Gillberg, 2009, cited in $n = 22$ studies, only three of which used a Swedish RMET). Notably, while the sample in Hallerbäck et al. (2009) consisted of university students, they administered a Swedish translation of the 28-item child version of the RMET, and this study was therefore excluded from the current review.

We also noticed that five studies that used the English language RMET cited Olderbak et al. (2015) as evidence for test-retest reliability. This study administered the English language version of the RMET. However, it did not assess test-retest reliability directly, but rather cited

⁷ Concerns have been raised about an overreliance on and misuse of quantitative validity evidence and cut-off values in validation research (Maul, 2017; McNeish & Wolf, 2023), and we agree there are problems with current construct validation practices. Nonetheless, because a primary aim of the current review is to evaluate current validity evidence reporting practices in the RMET literature, we compare the existing validity evidence for RMET scores with cut-off values that remain widely cited.

⁸ In the pre-registration, we stipulated that we would only extract validity information from a subset of studies if we identified over 1000. However, because so few studies provided validity evidence, we deviated from our pre-registration and extracted validity evidence from all included studies.

⁹ While Cohen's Kappa is frequently reported for inter-rater reliability, it assumes that the number of responses in each category are relatively equal. Because the categories are very unequal in our study, Kappa would be negatively biased, and it is not an appropriate representation of inter-rater reliability.

¹⁰ With the exception of Fig. 5, studies that were coded as calculating evidence and citing evidence from previous research are included in both the total percentage of studies that “calculated evidence” and “cited evidence from previous research” (i.e., they are counted twice). Fig. 5 reports the percentage of studies that did both.

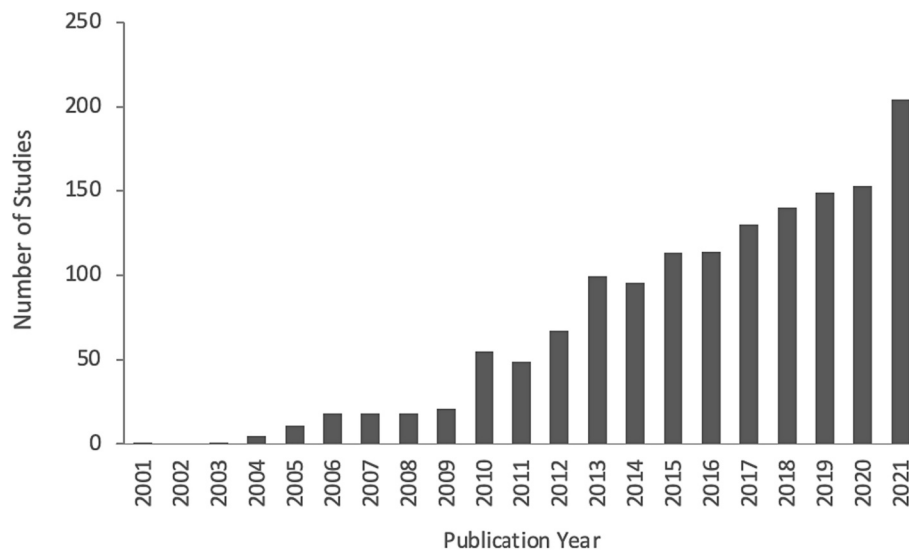


Fig. 2. Number of Publications per Year Reporting Multi Subject Empirical RMET Studies. *Note.* The apparent jump in the number of publications in 2021 could be due, at least in part, to our inclusion of studies that were published “online first” in 2021, which may subsequently have been assigned a publication year of 2022 (see footnote 5).

test-retest reliability from previous studies that did not administer the English language version.

2.3.2. Internal consistency

Internal consistency was calculated in 10.7% ($n = 157$) of studies and 8.6% ($n = 126$) cited evidence from previous research, while 84.5% ($n = 1234$)¹¹ did not provide any evidence. Of those studies that reported a metric for internal consistency, 86.6% ($n = 136$) calculated Cronbach's alpha (α). Other statistics were reported much less frequently and included Kuder-Richardson 20 (6.7%, $n = 10$),¹² split-half reliability (5.7%, $n = 9$), omega estimates (3.2%, $n = 5$), Rasch estimates (0.1%, $n = 1$), and fit statistics for confirmatory factor analysis of a single factor model (0.1%, $n = 1$). The most frequently cited source for the internal consistency of RMET scores was Vellante et al.'s (2013) validation study of an Italian version of the RMET, which was referenced in 46.8% of studies that cited previous evidence ($n = 59$, only 18 of which administered an Italian RMET).

Calculated levels of internal consistency for RMET scores were mixed. We extracted 164 alpha point estimates from 132 studies,¹³ with values ranging from 0.17 to 0.96, and 57.9% of these values fell below the conventional minimum acceptable value of 0.70 (see Fig. 4). We also extracted 18 omega point estimates from five studies, with values ranging from 0.43 to 0.85 (61.1% were below 0.70), ten Kuder-Richardson 20 point estimates from nine studies, ranging from 0.45 to 0.71 (90% were below 0.70), and nine split-half reliability point estimates from eight studies, ranging from 0.50 to 0.75 (55.6% were below 0.70). For 48.2% ($n = 93$) of the provided estimates, no interpretation or evaluation of the estimate was provided. For 23.8% ($n = 46$) of estimates, the metric was interpreted as indicating an acceptable level of

internal consistency. For 24.9% ($n = 48$) of estimates, it was noted that the values were low, but of these, only 15 studies mentioned that this might limit the interpretability of the RMET scores. Nine studies stated that the low value was not unsurprising due to some characteristic of the RMET (e.g., variability in the difficulty of items or the dichotomous scoring method for each item), and five studies noted that the values were in line with previous studies. Two studies provided other sources of evidence to support the interpretation of RMET scores despite the low estimates of internal consistency.

2.3.3. Factor structure

Factor analysis was only conducted on RMET scores in 0.9% ($n = 13$) of studies and 1.0% ($n = 14$) of studies cited results from previous research, while 98.6% ($n = 1440$) of studies did not provide any information on factor structure. In the identified studies, factor structure was assessed via exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and multi-group CFA (which is a type of CFA that simultaneously models the factor structure of test scores for multiple groups). Nine of the 13 studies that included factor analysis conducted CFA only; one conducted multigroup CFA only; and three conducted both EFA and CFA.

One of the three studies that conducted EFA used EFA as a method of selecting items for an abbreviated RMET based on factor loadings on a single factor model (Chander et al., 2020). As such, they did not discuss model fit. However, they did report the factor loadings for each item, and only 20 out of the 36 items had factor loadings >0.3 . They also conducted CFA on abbreviated versions of the RMET in order to identify a short-form version for use in older adults. The other two studies that conducted EFA failed to identify a well-fitting factor model. First, administering the English language version, Olderbak et al. (2015) found that the best model fit was a five-factor model; however, they ultimately rejected this model, in part because nine items (25%) failed to load sufficiently onto any of the five factors. Additionally, they noted that the identified factors lacked any obvious conceptual connection (e.g., dividing items according to the emotional valence of the target response or the sex of the person in the image). Second, administering a Serbian language version of the RMET, Dordevic et al. (2017) identified a two-factor model, which they also rejected due to poor model fit and the lack of a plausible conceptual account for the two factors.

The CFAs on RMET scores spanned six language versions and included models for both the full 36-item RMET and abbreviated

¹¹ Fernandez-Abascal et al. (2013) and Lindeman, Svedholm-Häkkinen, and Lipsanen (2015) claimed that the nature of the RMET makes it difficult or inappropriate to calculate alpha, yet they did not provide an alternative, more appropriate statistic.

¹² One study used Kuder Richardson 21, a variation of Kuder-Richardson 20 that assumes all items are of a similar level of difficulty.

¹³ These statistics include all reported alpha values, which span clinical and non-clinical groups, and translated and modified RMET versions. They exclude four studies that reported ranges of alpha values [>0.70 Burke et al., 2020; Cucchi, Hampton, & Moulton-Perkins, 2018], > 0.75 [Carroll, Montrose, & Burke, 2021], and 0.60 to 0.77 [Li et al., 2020]

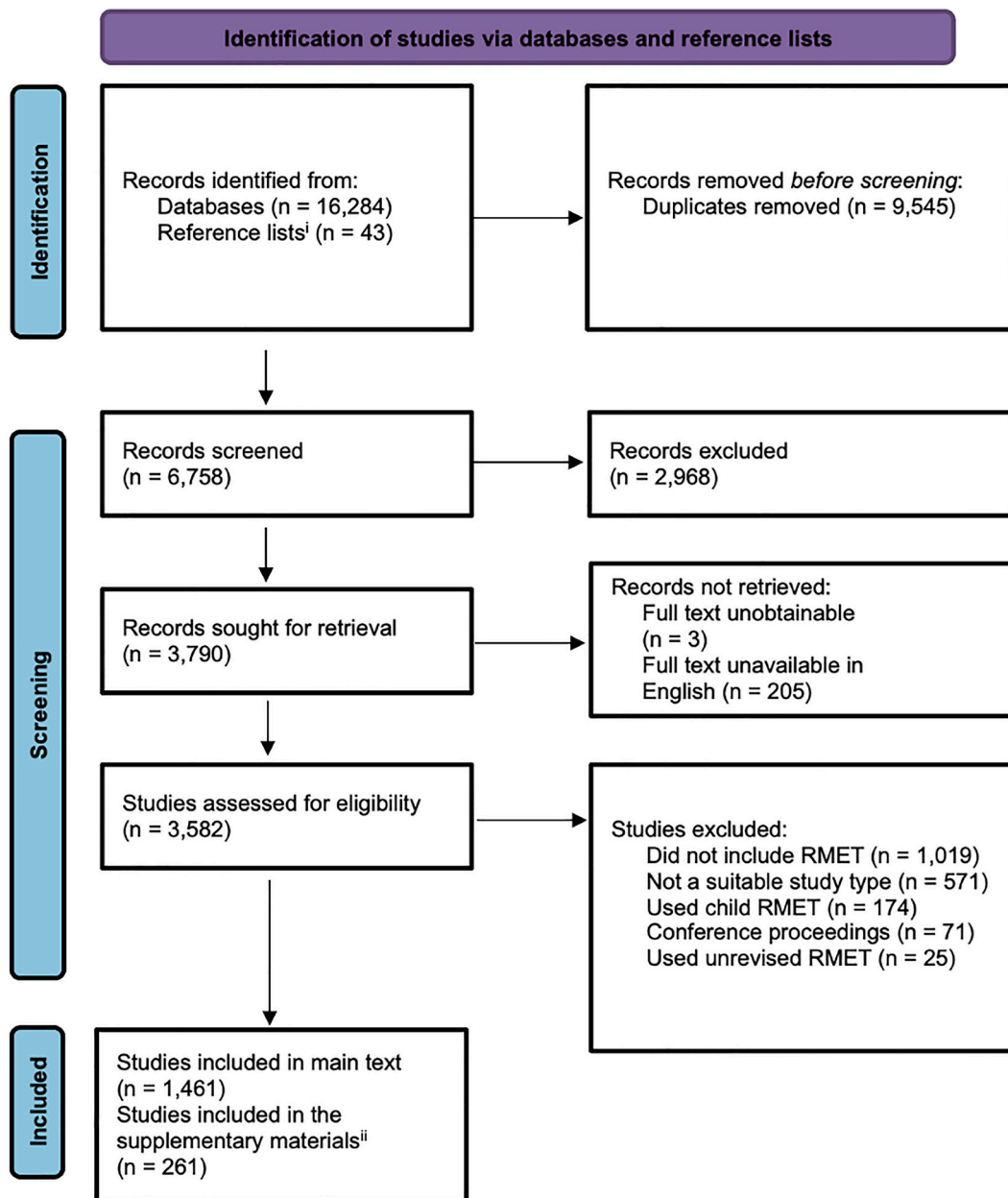


Fig. 3. PRISMA Flow Diagram of the Study Selection Process. *Note.* ⁱReference lists were not systematically searched. However, additional studies were identified when cited as sources of validity evidence for RMET scores or included in systematic reviews as a study that administered the RMET. ⁱⁱThe main text reports our findings for multi-subject empirical studies, where a “study” refers to a publication that reports findings from one or more empirical studies that used the RMET. Supplementary materials section S3 summarises validity evidence for single subject empirical studies, systematic and scoping reviews, meta-analyses, and protocols for randomised control trials.

versions. With the exception of [Chander et al. \(2020\)](#), all of the 13 studies that conducted CFA tested at least one factor model based on the 36-item RMET. We focus on the results from the factor analyses of the full 36-item test as this is the most widely used version. The fit statistics for these models are reported in Table S5. The RMET is generally interpreted as a unidimensional test producing a single sum score, and the most frequently tested models were single factor models ($n = 10$). The second most frequently tested models were three-factor models based on the valence subscores proposed by [Harkness et al. \(2005, n = 6\)](#). These subscores divide items into positive, negative, and neutral categories.

When testing factor models, studies predominantly evaluated models based solely on global model fit as assessed with fit indices. The values reported for the fit indices were mixed for both single and three-factor

models. In general, incremental fit indices (comparative fit index [CFI] and Tucker-Lewis index [TLI]) and the χ^2 test indicated poor model fit, while the other relative fit indices (standardised root mean square residual [SRMR] and root mean square error of approximation [RMSEA]) indicated acceptable model fit. With the exception of [Redondo and Herrero-Fernández \(2018\)](#), who reported an acceptable CFI value and unacceptable RMSEA value, this pattern held across language versions. There was variation in how researchers interpreted the mixed fit indices, with some concluding that model fit was good overall despite the lower values for some fit indices, while others concluded that model fit was poor. Overall, 54.5% ($n = 6$) of the 11 studies that conducted CFA using the 36-item RMET interpreted their results as indicating acceptable model fit for at least one tested model (see Table S5).

Only six studies reported factor loadings for individual test items,

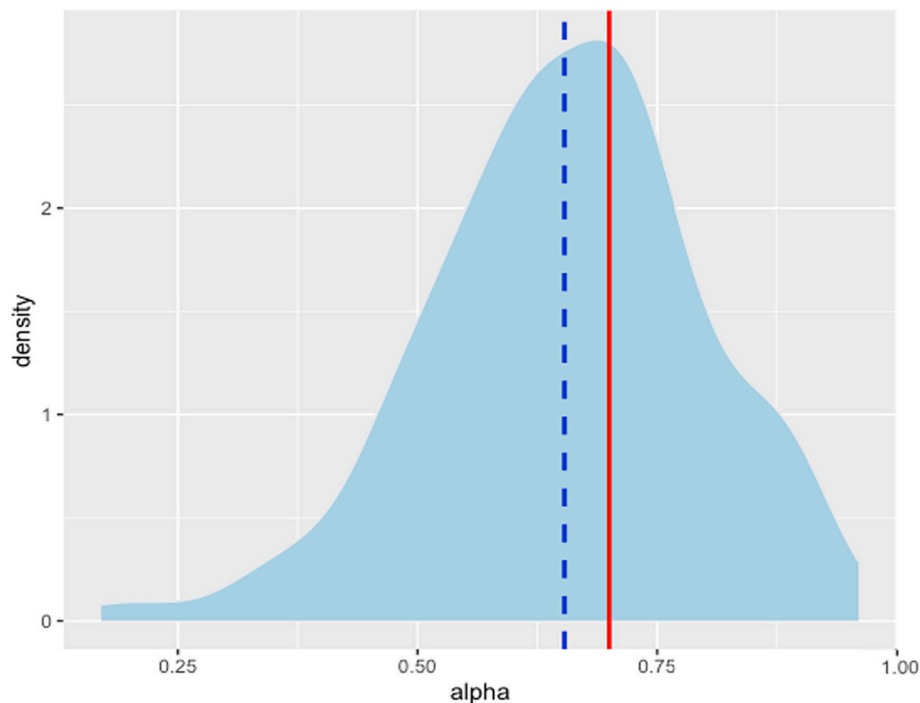


Fig. 4. Density Plot of Alpha Point Estimates of Internal Reliability for RMET Scores, *Note.* Alpha point estimates ($n = 164$) are based on samples from both clinical and non-clinical groups and span multiple language versions and modified versions of the RMET. The solid line represents the widely accepted alpha cut-off of 0.70. The dashed line shows the mean alpha point estimate of 0.65.

despite factor loadings being integral to model fit evaluation and important to report (Kline, 2023). Notably, it has been shown that the two fit indices that frequently indicated good model fit in the RMET literature (i.e., RMSEA and SRMR), improve as factor loadings decrease (lower factor loadings indicate lower item reliability), an effect referred to as the “reliability paradox” (Hancock & Mueller, 2011; McNeish & Wolf, 2023). Consistent with this paradox, Savalei (2012) demonstrated through simulation that RMSEA can indicate acceptable fit for a single factor model even if there are multiple latent factors (i.e., the correct model is multidimensional) when factor loadings are below 0.5. Where reported, the factor loadings for RMET items were frequently below 0.5. In fact, in three of the four studies that reported the factor loadings for all 36 RMET items (Dordevic et al., 2017; Olderbak et al., 2015; Vellante et al., 2013; cf. Petrocchi, Filipponi, Antonietti, Levante, & Lecciso, 2021), *only one* RMET item had a factor loading above 0.5, while at least 20 items had loadings below the recommended minimum of 0.3. Two additional studies (Redondo & Herrero-Fernández, 2018; Topić & Kovačević, 2019) removed items with low factor loadings but only reported the factor loadings for the retained items. Redondo and Herrero-Fernández (2018) retained 19 items, two of which had loadings above 0.5, while Topić and Kovačević (2019) retained seven items that all had loadings above 0.5. In contrast, Petrocchi et al. (2021) identified 21 items with loadings above 0.5 and also stated that “CFA of the 36 RMET items confirmed its single-factor structure” (p. 566); however, this study only reported statistics for a shortened version of the RMET with three items removed because their factor loadings were not statistically significant.

In addition to assessing global model fit using fit indices, a matrix of residuals or a description of the pattern of residuals should always be reported for factor analyses to provide information about local model fit (Kline, 2023). An examination of local model fit can help identify areas of model misfit, which can exist even when fit indices indicate good global model fit. Given the mixed findings for the global fit indices in the RMET literature, an examination of local fit could be extremely informative. However, none of the studies provided a matrix of residuals, although one study (Olderbak et al., 2015) did describe the pattern of

the residuals, noting that there was large residual variances (i.e., variance not explained by the model) for the majority of items.

Other analyses of the structural properties of the RMET presented in studies included a multi-group factor analysis of the relationship between race and performance on the English RMET, which reported significant differences in the psychometric properties of RMET scores for participants based on race and home language (Van Staden and Callaghan, 2021), and two item response theory analyses, with conflicting results (Black, 2019; Preti, Vellante, & Petretto, 2017). Black (2019) found best fit for a 2PL model, while Preti et al. (2017) found best fit for the Rasch (1PL) model.

2.3.4. Convergent, discriminant, and known group validity evidence

In 10.7% ($n = 157$) of studies, convergent validity evidence was calculated¹⁴ using test scores from tests administered within the study and 7.8% ($n = 114$) of studies cited evidence from previous research, while 83.9% ($n = 1226$) of studies did not provide any convergent validity evidence. Only 2.9% ($n = 42$) of studies calculated discriminant validity evidence based on tests administered within the study, while 3.5% ($n = 51$) cited evidence from previous research, and 94.9% ($n = 1387$) did not provide any discriminant validity evidence. The RMET has been used to measure a variety of social cognitive abilities (see “Variability in What the RMET is Used to Measure” section below), and in line with this, convergent validity was assessed against a variety of measures of ToM, empathy, emotion recognition, and autistic traits, while discriminant validity was *also* assessed against measures of empathy and emotion recognition. More straightforwardly, discriminant validity of the RMET was assessed against tests of cognitive abilities outside of social cognition, including IQ, verbal ability, and executive functioning.

In 3.2% ($n = 47$) of studies known group validity evidence was

¹⁴ As discussed in the methods section, we coded convergent, discriminant, and known validity as a “calculated” if correlations between measures or groups were discussed in relation to what the RMET measures, even if these types of validity evidence were not explicitly mentioned.

calculated based on reported participant data and 11.4% ($n = 166$) of studies cited evidence from previous research, while 87.6% ($n = 1280$) of studies did not provide any known group validity evidence. Given Baron-Cohen's (1995) influential theory arguing that a specific ToM deficit is a key underlying factor in autism, it is unsurprising that autism was by far the most frequently reported "known group", appearing as a source of validity evidence in 151 studies. It has also been argued that deficits in social cognitive abilities play a key role in functional outcomes in schizophrenia (Green, Horan, & Lee, 2015; Pinkham et al., 2016; Weinreb, Li, & Kurtz, 2022), motivating the interpretation of schizophrenia as a source of "known group" validity evidence in 29 studies. Males are also interpreted as "known group" validity evidence based on the extreme male brain and empathising-systematising theories of autism proposed by Baron-Cohen (Baron-Cohen, 2002, 2009; Baron-Cohen et al., 2015), which predict that, on average, females have higher levels of empathy than males, and thus should perform better on measures of social cognitive ability. The lower performance of males relative to females was identified as a "known group" in 19 studies. Other studies discussed the ability of the RMET to differentiate clinical from non-clinical groups without specifically identifying a clinical population; for example, "adults with atypical development" (Fastame & Carta, 2021, p. 2) and "adults with clinically meaningful outcomes" (Raffo De Ferrari et al., 2015, p. 510). For complete details of known groups, including quotations of the information on which coding decisions were based see the "details of known groups" sheet in the "External validity evidence" spreadsheet on the OSF page.

2.4. Reported strengths and limitations of RMET scores

While extracting validity evidence, we also recorded when studies highlighted what they took to be strengths or limitations of the RMET. Some of these strengths and limitations are related to the sources of

validity evidence that we coded; however, they also capture a broader range of characteristics of the RMET and provide considerable insight into how researchers tend to view the usefulness of the RMET. The strengths we most frequently identified include its widespread use and its being a valid and/or reliable measure. Notably, the RMET was explicitly referred to as a "valid" or "validated" measure in 15.2% of studies ($n = 222$; see Table S6 for illustrative quotes). The most frequently identified limitations include reliance on verbal ability, uncertainty about what the test measures, poor ecological validity, and poor internal consistency (see Tables S7 and S8 and the "Strengths and limitations" spreadsheet on the OSF page for further details of the identified strengths and limitations).

2.5. Validity evidence summary

Fig. 5 shows the percentage of studies that provided evidence for each of the six categories of validity evidence that we coded. We found that 63.3% ($n = 925$) of studies did not calculate or cite any of these types of validity evidence, while 13.6% ($n = 198$) of studies cited at least one type of validity evidence but did not calculate any new validity evidence, 11.6% ($n = 169$) of studies calculated at least one type of validity evidence, but did not cite any evidence from previous research, and 11.6% ($n = 169$) of studies calculated and cited evidence at least one type of validity evidence (the calculation and citation could have been from different categories of validity evidence).

2.6. Modifications

Modifications to the RMET itself included increasing (0.2%, $n = 3$) or decreasing (2.2%, $n = 32$) the number of mental state descriptors, increasing (1.4%, $n = 21$) or decreasing (7.6%, $n = 111$) the total number of test items, and limiting response times (3.6%, $n = 53$).

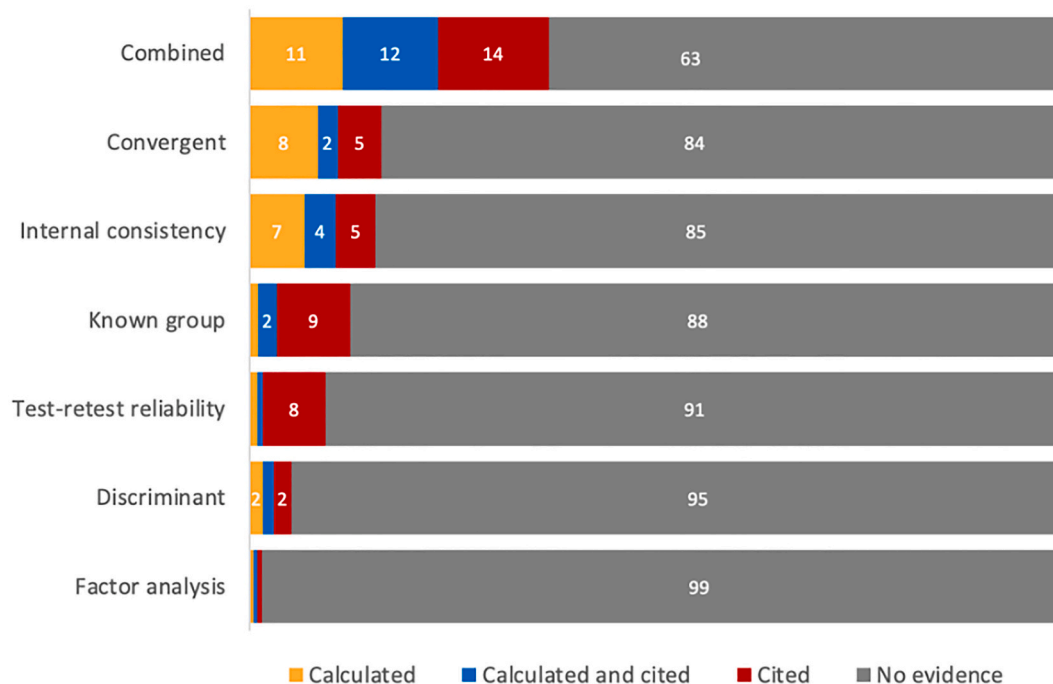


Fig. 5. Percentage of Studies Reporting Validity Evidence, *Note.* "Calculated" = new validity evidence was calculated using participant data (but evidence was not cited from previous research); "Cited" = validity evidence was cited from previous research (but new validity evidence was not calculated using participant data); "Calculated and cited" = new validity evidence was calculated using participant data and validity evidence was cited from previous research; "No evidence" = no validity evidence was provided. The "Combined" bar shows the percentage of studies that provided at least one of the coded categories of validity evidence. Studies are coded as "Calculated and cited" even if the new validity evidence calculated using participant data and the validity evidence cited from previous research were provided for different categories of validity evidence. For example, a study would get this coding if it calculated internal consistency using participant data and cited previous evidence of test-retest reliability.

Modifications to RMET scoring included calculating subscores (e.g., item difficulty, valence; 11.8%, $n = 172$), calculating a composite score with other measures of social cognition (1.6%, $n = 24$), and calculating z-scores (1.6%, $n = 23$). Despite the original version of the RMET being untimed, 4.2% ($n = 62$) of studies analysed response times. While this does not directly impact RMET scores, the interpretation of response times should be supported with validity evidence. Details of additional modifications can be found in the “Modifications” spreadsheet on the study’s OSF page. In total, we identified modifications in 28.7% ($n = 420$) of studies. The use of translated versions (reported below), which often had to be inferred, was coded separately. However, the percentage of studies with modifications increases to 61.4% ($n = 897$) when including studies where it was stated or inferred that a translated version of the RMET was administered.

2.6.1. Subscores

Subscores were analysed in 11.8% ($n = 172$) of studies. While the original scoring method produces a single sum score based on the total number of correct responses, studies also analysed subscores based on valence (i.e., positive, negative, neutral items; 9.9%, $n = 144$), item difficulty (1.8%, $n = 26$), sex of the pictured individual (1.0%, $n = 14$), and the type of mental state (e.g., cognitive vs emotional/social mental states, 0.5%, $n = 7$). We discuss valence and item difficulty subscores in more detail below.

The method for subdividing the RMET into valence categories for calculating valence subscores varied across studies. Some subscales were based on two valence categories (negative and positive, e.g., Macchia et al., 2021; Zamoscik et al., 2021), while other subscales included a third, neutral category (e.g., Harkness et al., 2005; Scott, Levy, Adams Jr, & Stevenson, 2011). There was also variation in how the valence of each item was determined, with some studies using ratings of the target response presented in isolation (e.g., Hezel & McNally, 2014), others using ratings of the photo of the eyes presented in isolation (e.g., Scott et al., 2011), and others using ratings of the photo presented with the target response (e.g., Harkness et al., 2005).

The subdivision into positive, negative, and neutral categories introduced in Harkness et al. (2005), is the most widely used valence categorisation (Hudson, Shamblaw, Harkness, & Sabbagh, 2020). Harkness and colleagues allocated items to valence categories based on the ratings of 12 female undergraduate students who rated each photo paired with its target response. They did not report any validity evidence specific to the subscores. However, they did note that the level of difficulty varied significantly between the different valence categories.

Validity evidence for valence subscores in subsequent research is also very limited. Of the studies that calculated valence subscores, only one (da Costa, Vrabell, Zeigler-Hill, & Vonk, 2018) reported internal consistency of the valence subscores separately. Using Harkness and colleagues’ valence categories, the reported alpha values for two of the three subscales were well below the values generally interpreted as acceptable (positive items 0.45, negative items 0.51, neutral items 0.68). As discussed above, six studies reported results from CFA of three factor valence-based models for the RMET (see Table S5). However, only one study (Lee, Nam, & Hur, 2020) concluded that there was acceptable model fit, and this claim was made despite the value for the reported relative fit index being very low (CFI = 0.45; compare with recommended value of > 0.95 ; Furr, 2022) and factor loadings and residuals not being reported. There was no validity evidence specific to subscores reported for any of the other six categories of validity evidence that we coded. However, Hudson et al. (2020, p. 623) highlighted potential issues with the validity of valence subscores. Based on valence ratings from 164 participants, they concluded that “valence categories are essentially arbitrary and largely influenced by sample size.” Additionally, using an existing dataset, they demonstrated that analyses based on the different valence subscales that have been reported in the literature result in contradictory conclusions.

Difficulty subscores were calculated in 1.8% ($n = 26$) of studies. As

with valence subscales, different methods were used by different researchers to divide RMET items into difficulty categories, with some researchers using the correct response rates in pilot studies (e.g., Domes, Heinrichs, Michel, Berger, & Herpertz, 2007) while others used response rates from previously published studies (e.g., Megias-Robles et al., 2020). No validity evidence specific to difficulty subscores was reported for any of the types of validity evidence we coded.

2.6.2. Abbreviated versions

The number of RMET items was reduced in 7.6% ($n = 111$) of studies.¹⁵ Reasons for abbreviating the test included reducing administration time (e.g., Chander et al., 2020; Hassenstab, Dziobek, Rogers, Wolf, & Convit, 2007) and improving the test’s psychometric properties by removing problematic items (e.g., Olderbak et al., 2015; Peterson & Miller, 2012). Methods of determining which items to retain in abbreviated versions also varied. Some researchers used statistically driven methods (e.g., Hassenstab et al., 2007; Olderbak et al., 2015), while other researchers arbitrarily selected items based on their position in the original test. For example, Bora, Eryavuz, Kayahan, Sungu, and Veznedaroglu (2006) selected the first 27 items, while Farago, Holmén, Holzmeister, Kirchler, and Razen (2022) used every other item starting with the first item. Of the studies that administered an abbreviated version of the RMET, 22.5% ($n = 25$), calculated internal consistency evidence, 4.5% ($n = 5$) calculated test-retest reliability evidence, 4.5% ($n = 5$) calculated factor structure evidence, 17.1% ($n = 19$) calculated convergent validity evidence, 3.6% ($n = 4$) calculated discriminant validity evidence, and 6.3% ($n = 7$) calculated known group validity evidence.

In 2.1% ($n = 31$) of studies, the number of mental state descriptors for each RMET item was reduced. In five of these studies, the words were removed altogether with participants asked to rate the valence of the mental state expressed in the pair of eyes (Kynast & Schroeter, 2018; Meyer & Morey, 2015), match the mental state expressed in the pair of eyes to the mental state expressed in another pair of eyes (Van den Stock et al., 2017), generate their own label for the emotion expressed in the eyes (Betz, Hoemann, & Barrett, 2019), or to think about the mental state being expressed in the eyes while lying in an fMRI brain scanner (Platek, Keenan, Gallup, & Mohamed, 2004). It was typically unspecified how it was determined which foil(s) was retained. However, Riem, Bakermans-Kranenburg, Voorthuis, and van Ijzendoorn (2014) included a pilot study and retained only the foil that was selected the least often. Only two studies calculated new validity evidence for the modified version. Meyer and Morey (2015), who removed all the mental state choices and had participants rate whether the eyes were expressing a negative emotion or not, reported an alpha value of 0.64. They suggested that the reliability of their modified version was similar to the original version, given the low alpha levels reported elsewhere. Another study provided convergent validity evidence based on a positive correlation with another measure of social cognition (Gleichgerricht, Torralva, Roca, Pose, & Manes, 2011).

¹⁵ Some translated versions of the RMET contain fewer than 36 items because items were removed during the validation process, usually due to low levels of participants picking the “correct” response or inadequate psychometric properties (e.g., Yildirim et al., 2011). We did not code studies using these translated versions as using an abbreviated version of the RMET. However, because studies did not always explicitly identify which language version of the RMET they had used, some of these translated versions may have been coded as abbreviated.

2.6.3. Translations

In total, we identified or inferred 39 different language versions of the RMET (see Table S9).¹⁶ Because not all studies explicitly identified which language version of the RMET was administered, we coded language as either “explicit”, “inferred”, or “unclear”. The explicit coding was used in instances where (1) the language version was explicitly stated in the paper, (2) acknowledgment was given to someone for providing a specific translated version, or (3) a citation for a validation study of a translated version was provided immediately after the RMET was introduced in the methods section provided the language version was consistent with demographic details provided for the participants. The “unclear” coding was used when the RMET was administered in a country with multiple widely used languages or the affiliated institutions were from countries with a variety of different languages.

In 25.1% ($n = 367$) of studies the language of the RMET was explicitly identified. In these cases, we also coded whether a citation was provided for translated versions. Of the 351 studies that explicitly identified using a translated version, 67.8% ($n = 238$) provided a citation for the translated version, 3.7% ($n = 13$) only stated that they downloaded the test from Autism Research Centre’s website (this is the website of the lab that created the English language RMET and provides translated versions for 28 languages¹⁷), 7.4% ($n = 27$) stated that they translated the RMET but did not identify themselves as an RMET validation study, 5.4% ($n = 19$) stated that one of their primary aims was to validate a translation of the RMET, 2.0% ($n = 7$) acknowledged someone else for providing a translated version, and 13.4% ($n = 47$) provided no information about the source of the translation.

In 74.8% ($n = 1093$) of studies, the language version was not explicitly identified, and we inferred the language from information in the paper. Inference criteria included an explicit statement of the language of another measure ($n = 147$, $n = 140$ where inferred language not English), participant language ($n = 118$, $n = 32$ where inferred language was not English), language-based exclusion/inclusion criterion ($n = 83$, $n = 12$ where inferred language was not English), location of recruitment or data collection ($n = 204$, $n = 103$ where inferred language was not English), and location of affiliated institutions ($n = 350$, $n = 45$ where inferred language was not English). Full details can be found on the “Languages” spreadsheet on the study’s OSF page.

In 62.7% ($n = 685$) of the studies in which the language version was not explicitly identified, we inferred that only the English language version was administered. In these cases, we would not expect an explicit language statement given that the RMET was originally published in English. However, in 33.6% ($n = 367$) of the studies where language was not explicitly identified, we inferred that the RMET was administered in at least one language other than English. Combined with studies that explicitly stated that they administered a translated version, this indicates that approximately half of the total number of studies in this review (49.1%, $n = 718$) administered a translated version of the RMET. A reasonable inference of language was not possible in 2.9% ($n = 43$) of the studies, and they were coded as “unclear”.

2.7. Constructs the RMET is used to measure

We identified considerable variability in the terminology used to describe what the RMET measures. Some of the terms we identified were used to describe what the RMET measures in the initial validation study (“theory of mind”, $n = 541$; “mentalising”, $n = 130$; “mind-reading”, $n = 51$; “social sensitivity”, $n = 38$; “social intelligence”, $n = 24$) or in later

studies by Baron-Cohen and colleagues (“affective ToM”, $n = 152$; “cognitive empathy”, $n = 76$; Baron-Cohen et al., 2015). Along with empathy ($n = 84$) and emotion recognition ($n = 206$), these terms remain among the most frequently used. However, we identified more than 50 unique ways of describing what the RMET measures (see Table S10 and the “Constructs” spreadsheet on the study’s OSF page for details of all terms).

2.8. Populations in which the RMET is used

The RMET was originally evaluated as a measure of social cognitive ability in autistic participants and members of the general public. However, we identified that its use has subsequently been extended to numerous other clinical and non-clinical populations. In 55.0% ($n = 804$) of the studies, a sample from one or more clinical groups was included. In total, the RMET has been administered to samples from more than 75 different clinical populations. The 10 most frequently sampled are listed in Table 1. Studies also included more than 50 non-clinical samples drawn from specific subsets of the general public such as Christians, inmates, twins, or managers (for details of additional clinical and non-clinical groups, see Table S11 and the “Sampled populations” spreadsheet on the study’s OSF page).

3. Discussion

The most striking finding from the present review of 1461 studies that used the RMET is the inadequacy of validity evidence reporting practices. Sixty-three percent of studies did not calculate new construct validity evidence for RMET scores or cite evidence from previous research for *any* of the six categories of construct validity evidence that we coded. Only 23% of studies calculated new validity evidence for *at least one* category of evidence despite the fact that 29% of studies used a modified scoring method or modified version of the RMET and at least 75 different clinical groups were investigated using the RMET. Moreover, in those cases where validity evidence was reported, RMET scores frequently failed to meet criteria that are conventionally taken to indicate acceptable levels of validity. These findings demonstrate that even a very widely used measure can lack validity evidence.

Only 12% of studies reported new evidence for the structural validity of RMET scores, and with the exception of test-retest reliability, the evidence was frequently weak. Fifty-eight percent of the reported alpha values (which was by far the most frequently reported evidence of internal consistency) were below the conventionally accepted minimum value of 0.70, often *well below*. The frequency of unacceptably low alpha values is especially surprising given the length of the RMET. As noted by Clark and Watson (2019, p. 1421): “with a large number of items, it is exceedingly difficult to avoid having a high reliability estimate, so coefficient alpha is virtually useless for scales containing 40 or more items.” Yet, despite the RMET being a long test with 36 items, alpha values for RMET scores were frequently low.

To date, the factor structure of RMET scores across studies remains unclear. Of the 13 studies in our review that conducted CFA, *only one*

Table 1
Most frequently sampled clinical populations.

Clinical group	Number of studies
Schizophrenia spectrum and psychosis disorders	254
Autism	147
Bipolar disorder	49
Depressive disorders	37
Parkinson’s disease	32
Eating disorders	27
Multiple sclerosis	28
Substance use disorders	26
Borderline personality disorder	24
Dementia	24

¹⁶ In some translated versions the photos of White eyes were replaced to make the test more appropriate for cross-cultural use. We included these versions in the scoping review, and we did not systematically code this as a feature of these studies.

¹⁷ The Autism Research Centre webpage does not provide any validity evidence specific to translated versions.

study identified a factor model for the unabbreviated RMET with factor loadings above the conventionally recommended minimum level of 0.30 for the majority of items and reported acceptable levels of model fit across key fit indices (Petrocchi et al., 2021). This study was not published until 2021, meaning that, to the best of our knowledge, the RMET was in use for more than two decades before the first comprehensive analysis of the factor structure of RMET scores consistent with a unidimensional model was published. Additionally, this study used an Italian translation of the RMET, and the most comprehensive confirmatory factor analysis we identified of the English language RMET reported poor unidimensional model fit (Olderbak et al., 2015).

The lack of clarity on the factor structure of RMET scores has two important implications. First, the RMET was designed to have its 36 items summed together into a single score (Baron-Cohen et al., 2001) and, consequently, the interpretation of the single sum score should be supported by factor modelling evidence consistent with the existence of a single latent factor that influences performance on all test items within a given dataset (McNeish & Wolf, 2020; Widaman & Revelle, 2023). Likewise, when subscores are calculated for the RMET, factor modelling should be conducted to ensure that the factor structure matches the proposed subscales. Second, knowledge of the RMET's factor structure is required to select appropriate estimates of internal consistency (Flora, 2020; Savalei & Reise, 2019). While alpha was by far the most frequently reported method for estimating internal consistency for RMET scores, alpha assumes unidimensionality and equivalent factor loadings (Flora, 2020; Widaman & Revelle, 2023). Given the lack of evidence for unidimensional models in the RMET literature, alpha is likely an inappropriate measure of the internal consistency of RMET scores in many cases. Alternative methods are available to estimate internal consistency in the case of multidimensionality, including the omega family of reliability estimates (Flora, 2020; Furr, 2022). However, assessment of the factor structure remains necessary because these are based on factor modelling and the use of an inappropriate factor model can result in biased reliability estimates (Savalei & Reise, 2019).

In the initial validation study, known group validity was used as a key source of evidence to establish the RMET as a measure of social cognitive ability (Baron-Cohen, Wheelwright, Hill, et al., 2001), and our review indicates that the ability of the RMET to identify autistic participants as having lower average scores than members of the general population is one of the most frequently cited sources of validity evidence; this was mentioned in 10% of studies. However, as noted earlier, the assumption that a core feature of autism is a specific ToM deficit has been contested (Deschrijver & Palmer, 2020; Gernsbacher & Yergeau, 2019; Milton et al., 2022; Silverman, 2022), and factors other than ToM ability have been proposed as potentially contributing to lower RMET scores in autistic individuals, including discomfort viewing eyes (Higgins, Ross, Langdon, & Polito, 2023), emotion recognition difficulties related to alexithymia (Oakley et al., 2016), and differences in the extent to which autistic people focus on the relationship between the four mental state terms versus the relationship between these terms and the image when identifying the "correct" response (Johnston et al., 2008).

Our review also revealed at least 50 different terms for describing what the RMET measures. This astonishingly large number is consistent with arguments that a lack of conceptual clarity in the social cognition literature has led to a proliferation of overlapping constructs (Hall & Schwartz, 2022; Quesque & Rossetti, 2020; Schaafsma, Pfaff, Spunt, & Adolphs, 2015; Schurz et al., 2021), resulting in the same terminology being used to describe distinct cognitive processes (i.e., the jingle fallacy) and different terminology being used to describe the same cognitive process (i.e., the jangle fallacy). In fact, no less than 43 distinct definitions of empathy have been identified in the literature (Cuff, Brown, Taylor, & Howat, 2016).

This lack of conceptual clarity poses key challenges for evaluating the construct validity of RMET scores because construct validity is an assessment of the extent to which a measurement can be interpreted as a measure of a *particular* latent construct (Hall & Schwartz, 2022).

Additionally, inconsistency concerning which specific aspect(s) of social cognitive ability the RMET is considered to measure makes it far from clear whether correlations between RMET scores and scores on measures of related constructs (e.g., emotion recognition or empathy) should be interpreted as sources of convergent versus discriminant validity evidence (Clark & Watson, 2019). For example, Lee et al. (2020) used the RMET as a measure of emotion perception and interpreted the lack of a statistically significant correlation with the cognitive empathy domain of the Interpersonal Reactivity Index (Davis, 1983) as indicative of discriminative validity, supporting the use of the RMET as a measure of emotion recognition. In contrast, Dehning et al. (2013) used the RMET as a measure of cognitive empathy and interpreted a statistically significant correlation between RMET performance and performance on the Balanced Emotional Empathy Scale (Mehrabian, 1996) as evidence supporting the use of the RMET as a measure of cognitive empathy.

Notably, depending on whether a correlation is interpreted as a source of convergent or discriminant validity evidence the *same* correlation between the same two variables could be taken as evidence for or against the validity of RMET scores. This highlights the importance of conceptual clarification of what construct a psychological test is designed to measure and of assessing validity in relation to the specified construct (Bringmann, Elmer, & Eronen, 2022). Additionally, low correlations between the RMET and other measures of social cognition were often interpreted as indicating that the measurements capture different aspects of social cognition (e.g., Chen et al., 2017; Nemeth, Peterfalvi, Czeh, Tenyi, & Simon, 2020; Raimo et al., 2017) rather than as evidence against the validity of either measurement.

Another striking finding of this review is the variety of populations the RMET has been used to study, including at least 75 different clinical populations. This raises another key validity concern because characteristics specific to different populations can impact RMET performance in ways that reduce the interpretability of test scores in terms of social cognitive ability within those groups. Consequently, it cannot be assumed that RMET scores have the same structural properties or are equally valid within samples from different populations (Clark & Watson, 2019). For example, it has been argued that the RMET has limited sensitivity at higher levels of social cognitive ability, reducing its utility in non-clinical populations (Black, 2019); and different structural properties have been identified for RMET scores across cultural groups (Van Staden & Callaghan, 2021). In relation to clinical populations, Foley et al. (2019) suggested that lower performance on ToM tasks (including the RMET) in Parkinson's disease results from impairments in inhibition rather than ToM ability per se, while Russell, Schmidt, Doherty, Young, and Tchanturia (2009) raised the possibility that the sexualised nature of some of the RMET items might result in lower scores by anorexic female participants. As such, the validity of RMET scores in samples from different populations should be comprehensively evaluated based on a variety of types of validity evidence (Flake et al., 2017). While we did not subdivide the validity evidence according to each specific population sampled, the overall low levels of validity evidence reported and the fact that the RMET was administered in more than 75 different clinical groups suggests that comprehensive assessment of the evidence for the validity of RMET scores does not exist for many of these groups. For example, test-retest reliability was only assessed in 22 studies, while factor structure was only assessed in 13 studies.

A final notable finding of this review is the frequency of modifications to the RMET without a careful examination of the impact this might have on the validity of test scores. Including translation, 61% of the studies administered a modified version of the RMET or used a modified scoring method. Because modifications can impact the validity of test scores, scores from modified versions, including translations, should be validated independently (AERA et al., 2014; Appelbaum et al., 2018; Clark & Watson, 2019). While studies specifically designed to validate scores from translated versions of the RMET reported validity evidence specific to the modified version, numerous studies failed to even identify the language version that was administered, as indicated

by the large number of studies for which we inferred a translated version was used. Validity evidence specific to other modified RMET versions and scoring was rarely reported. Overall, where reported, validity evidence for modified RMET versions often indicated inadequate validity.

3.1. Implications and recommendations

The lack of engagement with validity evidence revealed in our review is contrary to guidelines and recommendations for psychology research. In particular, the American Psychological Association's reporting standards recommend that, when possible, new sample specific construct validity evidence should always be calculated and reported, and when evidence is cited from previous research, the basic demographic characteristics of the samples in which the evidence was calculated should be reported and compared to the current study's samples (Appelbaum et al., 2018). As we have demonstrated, research in the RMET literature falls well below this recommended standard, with numerous studies failing to cite existing validity evidence let alone report new validity evidence. These findings suggest that, in the social cognition literature at least, the minimum level of validity evidence required for publication needs to be drastically increased, bringing it in line with recommended standards and practices (e.g., AERA et al., 2014; Appelbaum et al., 2018). This should help protect against the production of bodies of research that resemble the RMET literature in providing little evidence that the psychological instruments on which research findings are based measure what they are intended to measure.

Given the findings of this review, we strongly advise against the ongoing use of the RMET as a measure of social cognitive ability. Our review also raises challenging questions about how to engage with the existing RMET literature because it shows that most research findings based on RMET scores are unsubstantiated due to inadequate validity evidence. Excepting any cases where construct validity evidence for RMET scores can be retrospectively obtained (e.g., via the re-analysis of raw data from these studies), we advise that research findings based on unsubstantiated RMET scores stop being cited as evidence for psychological theories, stop being used to inform clinical diagnosis and practice, and stop being disseminated to the general public. Furthermore, we suggest that existing psychological theories and clinical guidelines that rely heavily on unsubstantiated RMET findings should be reassessed. Our recommendations may sound extreme. However, construct validity is fundamental to the production of valid scientific research, and it needs to be taken much more seriously by psychologists (AERA et al., 2014; Clark & Watson, 2019; Flake & Fried, 2020).

Construct validity concerns are clearly not unique to the RMET. As the psychological community reflects on its practices in the wake of the replication crisis (Nelson, Simmons, & Simonsohn, 2018; Nosek et al., 2022; Open Science Collaboration, 2015), the construct validity of psychological measurements has come under increased scrutiny (Flake & Fried, 2020). Problems with construct validation have been identified in high-ranking psychology journals (Flake et al., 2022) and across a variety of psychological research areas including anger and hostility (Vassar & Hale, 2009), depression (Fried et al., 2022), personality and social psychology (Flake et al., 2017), health education and behaviour (Barry, Chaney, Piazza-Gardner, & Chavarria, 2014), social cognition (Feest, 2020), and well-being (Alexandrova & Haybron, 2016). In particular, there is growing evidence that a large proportion of empirical studies include little to no evidence for key dimensions of construct validity, indicating that current validity evidence reporting practices are woefully inadequate (Flake et al., 2017; Flake et al., 2022; Flake & Fried, 2020; Fried et al., 2022; Hussey & Hughes, 2020; Schimmack, 2021; Shaw et al., 2020; Slaney, 2017; Vassar & Hale, 2009). This suggests that a lack of engagement with construct validity is a pervasive problem in psychological science. Additionally, our finding that there is a plurality of terms used to describe what the RMET measures provides a compelling example of how a lack of conceptual clarity can negatively impact the construct validation process (Alexandrova & Haybron, 2016;

Bringmann et al., 2022).

3.2. Limitations

Our review of validity evidence reporting practices in the RMET literature has several limitations that are worth highlighting. First, screening of studies was completed by a single individual and only a subset of the studies (15%) was double coded during data extraction. While the levels of inter-rater reliability for the studies that were double coded were high, it is possible that there are coding errors. Nonetheless, the dataset is so large and most patterns of findings are so pronounced that even if some validity evidence was missed or miscoded, the overall pattern of results is unlikely to change. Second, our review may underestimate the validity evidence reporting practices for translated versions of the RMET. We excluded 205 studies because the full text was not available in English, and these studies may include validity evidence for RMET scores. Additionally, our search strategy of identifying studies that cited Baron-Cohen and colleagues' initial validation study for the RMET may have missed studies that used translated versions and only provided a citation for the translated version. Third, we coded studies as having cited previous evidence even if they did not provide a citation to support the validity claim, and we did not differentiate studies that did and did not provide a citation. We also coded convergent, discriminant, and known group evidence even if these types of validity were not explicitly mentioned. This means that our review almost certainly overestimates the extent to which researchers in this literature genuinely engaged with construct validity evidence. Fourth, we limited our review to six categories of quantitative construct validity evidence, which are not exhaustive, and we did not review qualitative construct validity evidence (e.g., focus groups or cognitive interviews to assess how participants interpret test items) or theory-based validity evidence, which are important forms of validity evidence that are frequently neglected (Alexandrova & Haybron, 2016; Wolf, 2023). Nonetheless, when we did encounter other types of construct validity evidence or sources of information that could be relevant to construct validity during the data extraction process, they were recorded in "other evidence" and "notes" sections in the data extraction template, and the details can be found in the "Full dataset (all study types)" spreadsheets on the study's OSF page.

4. Conclusion

We used a scoping review methodology to examine validity evidence reporting practices in 1461 studies that administered the RMET (including 804 studies that included at least one clinical sample), focusing on six key types of validity evidence. We found that evidence to support the validity of RMET scores was infrequently reported. Strikingly, much of the validity evidence that was reported raised serious concerns about RMET scores being "valid". Despite this, we found that the RMET is widely reported to be a "validated" measure and that the number of studies published each year using the RMET is steadily increasing. Given our findings, we advise against further use of the RMET and urge considerable caution when interpreting existing findings based on RMET scores. Furthermore, our findings demonstrate that a scoping review can be a powerful tool for assessing the validity evidence for a widely used psychological measure, and they add important evidence to support recent claims that measurement practices in psychology are inadequate and that validity evidence reporting guidelines are routinely ignored. We hope that others will be inspired by our example to interrogate the validity of other psychological measurements.

Funding

WCH is funded by an Australian Government Research Training Program (RTP) Scholarship, a Macquarie University Research Excellence Scholarship, and a Macquarie University School of Psychological

Sciences Higher Degree Research Grant. RMR is funded by a Macquarie University Research Fellowship and the John Templeton Foundation (Grant ID: 62631). ED is funded by a Discovery Early Researcher Award (DECRA) from the Australian Research Council (ARC; Grant ID: DE220100087) and a BOF fellowship from Ghent University. WCH, DRK, ED, and RMR also received funding from the Macquarie University Ethics and Agency Research Centre.

Declaration of Competing Interest

None.

Acknowledgements

We would like to thank John Sutton for enthusiastically supporting this project during its early stages, Alexander Gillett and Richard Ramsey for their insightful feedback, and Isabelle Fayle for contributing to the data extraction and coding process.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cpr.2023.102378>.

References

- Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science*, 83(5), 1098–1109. <https://doi.org/10.1086/687941>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, 73(1), 3–25. <https://doi.org/10.1037/amp0000191>
- Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and “Reading the mind in the eyes”. *Intelligence*, 44(1), 78–92. <https://doi.org/10.1016/j.intell.2014.03.001>
- Baron-Cohen, B., D. C., Holt, R. J., Allison, C., Auyeung, B., Lombardo, M. V., ... Lai, M.-C. (2015). The “Reading the mind in the eyes” test: complete absence of typical sex difference in ~400 men and women with autism. *PLoS One*, 10(8), Article e0136521. <https://doi.org/10.1371/journal.pone.0136521>
- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MIT Press.
- Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Sciences*, 6(6), 248–254. [https://doi.org/10.1016/S1364-6613\(02\)01904-6](https://doi.org/10.1016/S1364-6613(02)01904-6)
- Baron-Cohen, S. (2009). Autism: The empathizing-systemizing (E-S) theory. *Annals of the New York Academy of Sciences*, 1156(1), 68–80. <https://doi.org/10.1111/j.1749-6632.2009.04467.x>
- Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, 38(7), 813–822. <https://doi.org/10.1111/j.1469-7610.1997.tb01599.x>
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the mind in the eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, 42(2), 241–251. <https://doi.org/10.1017/S0021963001006643>
- Baron-Cohen, S., Wheelwright, S., Spong, A., Scabill, V., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger syndrome. *Journal of Developmental and Learning Disorders*, 5(1), 47–78.
- Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior*, 41(1), 12–18. <https://doi.org/10.1177/1090198113483139>
- Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3), 303–308. <https://doi.org/10.1086/266520>
- Betz, N., Hoemann, K., & Barrett, L. F. (2019). Words are a context for mental inference. *Emotion*, 19(8), 1463–1477. <https://doi.org/10.1037/emo0000510>
- Black, J. E. (2019). An IRT analysis of the Reading the mind in the eyes test. *Journal of Personality Assessment*, 101(4), 425–433. <https://doi.org/10.1080/00223891.2018.1447946>
- Bora, E., Eryavuz, A., Kayahan, B., Sungu, G., & Veznedaroglu, B. (2006). Social functioning, theory of mind and neurocognition in outpatients with schizophrenia; mental state decoding may be a better predictor of social functioning than mental state reasoning. *Psychiatry Research*, 145(2–3), 95–103. <https://doi.org/10.1016/j.psychres.2005.11.003>
- Botha, M., Hanlon, J., & Williams, G. L. (2023). Does language matter? Identity-first versus person-first language use in autism research: A response to Vivanti. *Journal of Autism and Developmental Disorders*, 53(2), 870–878. <https://doi.org/10.1007/s10803-020-04858-w>
- Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science*, 31(4), 340–346. <https://doi.org/10.1177/09637214221096485>
- Burke, T., Pinto-Grau, M., Costello, E., Peelo, C., Loneragan, K., Heverin, M., ... Pender, N. (2020). The Reading the mind in the eyes test short form (A & B): Validation and outcomes in an amyotrophic lateral sclerosis cohort. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 21(5–6), 380–388. <https://doi.org/10.1080/21678421.2020.1772824>
- Carroll, G. A., Montrose, V. T., & Burke, T. (2021). Correlates of social cognition and psychopathic traits in a community-based sample of males. *Frontiers in Psychology*, 12, 656299. <https://doi.org/10.3389/fpsyg.2021.656299>
- Chander, R. J., Grainger, S. A., Crawford, J. D., Mather, K. A., Numbers, K., Cleary, R., ... Sachdev, P. S. (2020). Development of a short-form version of the reading the mind in the eyes test for assessing theory of mind in older adults. *International Journal of Geriatric Psychiatry*, 35(11), 1322–1330. <https://doi.org/10.1002/gps.5369>
- Chen, K. W., Lee, S. C., Chiang, H. Y., Syu, Y. C., Yu, X. X., & Hsieh, C. L. (2017). Psychometric properties of three measures assessing advanced theory of mind: Evidence from people with schizophrenia. *Psychiatry Research*, 257, 490–496. <https://doi.org/10.1016/j.psychres.2017.08.026>
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. <https://doi.org/10.1037/pas0000626>
- da Costa, H. P., Vrabel, J. K., Zeigler-Hill, V., & Vonk, J. (2018). DSM-5 pathological personality traits are associated with the ability to understand the emotional states of others. *Journal of Research in Personality*, 75, 1–11. <https://doi.org/10.1016/j.jrp.2018.05.001>
- Cucchi, A., Hampton, J. A., & Moulton-Perkins, A. (2018). Using the validated reflective functioning questionnaire to investigate mentalizing in individuals presenting with eating disorders with and without self-harm. *PeerJ*, 6, Article e5756. <https://doi.org/10.7717/peerj.5756>
- Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, 8(2), 144–153. <https://doi.org/10.1177/1754073914558466>
- Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126. <https://doi.org/10.1037/0022-3514.44.1.113>
- De Paoli, T., Fuller-Tyszkiewicz, M., Huang, C., & Krug, I. (2020). A network analysis of borderline personality disorder symptoms and disordered eating. *Journal of Clinical Psychology*, 76(4), 787–800. <https://doi.org/10.1002/jclp.22916>
- Dehning, S., Gasperi, S., Tesfaye, M., Girma, E., Meyer, S., Krah, W., ... Siebeck, M. (2013). Empathy without borders? Cross-cultural heart and mind-reading in first-year medical students. *Ethiopian Journal of Health Sciences*, 23(2), 113–122.
- Deschrijver, E., & Palmer, C. (2020). Reframing social cognition: Relational versus representational mentalizing. *Psychological Bulletin*, 146(11), 941–969. <https://doi.org/10.1037/bul0000302>
- Dodell-Feder, D., Ressler, K. J., & Germine, L. T. (2020). Social cognition or social class and culture? On the interpretation of differences in social cognitive performance. *Psychological Medicine*, 50(1), 133–145. <https://doi.org/10.1017/S00329171800404x>
- Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves “mind-reading” in humans. *Biological Psychiatry*, 61(6), 731–733. <https://doi.org/10.1016/j.biopsych.2006.07.015>
- Dordevic, J., Zivanovic, M., Pavlovic, A., Mihajlovic, G., Karlicic, I. S., & Pavlovic, D. (2017). Psychometric evaluation and validation of the Serbian version of “Reading the mind in the eyes” test. *Psihologija*, 50(4), 483–502. <https://doi.org/10.2298/PSI170504010D>
- Eddy, C. M. (2019). What do you have in mind? Measures to assess mental state reasoning in neuropsychiatric populations. *Frontiers in Psychiatry*, 10, 425. <https://doi.org/10.3389/fpsyg.2019.00425>
- Eddy, C. M., & Hansen, P. C. (2020). Predictors of performance on the reading the mind in the eyes test. *PLoS One*, 15(7), Article e0235529. <https://doi.org/10.1371/journal.pone.0235529>
- Farago, A., Holmén, M., Holzmeister, F., Kirchler, M., & Razen, M. (2022). Cognitive skills and economic preferences in the fund industry. *The Economic Journal*, 132 (645), 1737–1764. <https://doi.org/10.1093/ej/ueab092>
- Fastame, M. C., & Carta, V. (2021). Do cognitive reserve and executive functions matter to perform the reading the mind in the eyes test in late lifespan? *Applied Neuropsychology. Adult*, 1–9. <https://doi.org/10.1080/23279095.2021.1972422>
- Feest, U. (2020). Construct validity in psychological tests – The case of implicit social cognition. *European Journal for Philosophy of Science*, 10(1). <https://doi.org/10.1007/s13194-019-0270-8>
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-L](https://doi.org/10.1016/0895-4356(90)90158-L)
- Fernandez-Abascal, E. G., Cabello, R., Fernandez-Berrocal, P., & Baron-Cohen, S. (2013). Test-retest reliability of the “Reading the mind in the Eyes” test: A one-year follow-up study. *Molecular Autism*, 4(1), 33. <https://doi.org/10.1186/2040-2392-4-33>
- Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct validity and the validity of replication studies: A systematic review. *The American Psychologist*, 77(4), 576–588. <https://doi.org/10.1037/amp0001006>
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. <https://doi.org/10.1177/2515245920952393>

- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378. <https://doi.org/10.1177/1948550617693063>
- Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science*, 3(4), 484–501. <https://doi.org/10.1177/2515245920951747>
- Foley, J. A., Lancaster, C., Poznyak, E., Borejko, O., Niven, E., Foltynie, T., ... Cipolotti, L. (2019). Impairment in theory of mind in Parkinson's disease is explained by deficits in inhibition. *Parkinson's Disease*, 2019, 5480913–5480918. <https://doi.org/10.1155/2019/5480913>
- Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, 1(6), 358–368. <https://doi.org/10.1038/s44159-022-00050-2>
- Furr, R. M. (2022). *Psychometrics: An introduction* (4th ed.). SAGE.
- Gernsbacher, M. A., & Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Archives of Scientific Psychology*, 7(1), 102. <https://doi.org/10.1037/arc0000067>
- Gleichgerricht, E., Torralva, T., Roca, M., Pose, M., & Manes, F. (2011). The role of social cognition in moral judgment in frontotemporal dementia. *Social Neuroscience*, 6(2), 113–122. Article Pii 925719115 <https://doi.org/10.1080/17470919.2010.506751>
- Green, M. F., Horan, W. P., & Lee, J. (2015). Social cognition in schizophrenia. *Nature Reviews Neuroscience*, 16(10), 620–631. <https://doi.org/10.1038/nrn4005>
- Hall, J. A., & Schwartz, R. (2022). Empathy, an important but problematic concept. *The Journal of Social Psychology*, 162(1), 1–6. <https://doi.org/10.1080/00224545.2021.2004670>
- Hallerbäck, M. U., Lugnegård, T., Hjärthar, F., & Gillberg, C. (2009). The Reading the mind in the eyes test: Test-retest reliability of a Swedish version. *Cognitive Neuropsychiatry*, 14(2), 127–143. <https://doi.org/10.1080/13546800902901518>
- Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, 71(2), 306–324. <https://doi.org/10.1177/0013164410384856>
- Harkness, K. L., Sabbagh, M. A., Jacobson, J. A., Chowdrey, N. K., & Chen, T. (2005). Enhanced accuracy of mental state decoding in dysphoric college students. *Cognition & Emotion*, 19(7), 999–1025. <https://doi.org/10.1080/02699930541000110>
- Hassenstab, J., Dziobek, I., Rogers, K., Wolf, O. T., & Convit, A. (2007). Knowing what others know, feeling what others feel - A controlled study of empathy in psychotherapists. *Journal of Nervous and Mental Disease*, 195(4), 277–281. <https://doi.org/10.1097/01.nmd.0000253794.74540.2d>
- Herbet, G., Lafargue, G., Moritz-Gasser, S., Bonnetblanc, F., & Duffau, H. (2015). Interfering with the neural activity of mirror-related frontal areas impairs mentalistic inferences. *Brain Structure & Function*, 220(4), 2159–2169. <https://doi.org/10.1007/s00429-014-0777-x>
- Herbet, G., & Moritz-Gasser, S. (2019). Beyond language: Mapping cognition and emotion. *Neurosurgery Clinics of North America*, 30(1), 75–83. <https://doi.org/10.1016/j.nec.2018.08.004>
- Herbet, G., Rigaux-Viodé, O., & Moritz-Gasser, S. (2017). Peri- and intraoperative cognitive and language assessment for surgical resection in brain eloquent structures. *Neurochirurgie*, 63(3), 135–141. <https://doi.org/10.1016/j.neuchi.2016.10.011>
- Hezel, D. M., & McNally, R. J. (2014). Theory of mind impairments in social anxiety disorder. *Behavior Therapy*, 45(4), 530–540. <https://doi.org/10.1016/j.beth.2014.02.010>
- Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2023). The “Reading the Mind in the Eyes” test shows poor psychometric properties in a large, demographically representative U.S. sample. *Assessment*, 30(6), 1777–1789. <https://doi.org/10.1177/1073191122124342>
- Higgins, W. C., Ross, R. M., Polito, V., & Kaplan, D. M. (2023). Three threats to the validity of the Reading the mind in the eyes test: A commentary on Pavlova and Sokolov (2022). *Neuroscience and Biobehavioral Reviews*, 105088. <https://doi.org/10.1016/j.neubiorev.2023.105088>
- Hudson, C. C., Shablaway, A. L., Harkness, K. L., & Sabbagh, M. A. (2020). Valence in the Reading the mind in the eyes task. *Psychological Assessment*, 32(7), 623–634. <https://doi.org/10.1037/pas0000818>
- Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, 3(2), 166–184. <https://doi.org/10.1177/2515245919882903>
- Johnston, L., Miles, L., & McKinlay, A. (2008). A critical review of the eyes test as a measure of social-cognitive impairment. *Australian Journal of Psychology*, 60(3), 135–141. <https://doi.org/10.1080/00049530701449521>
- Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2022). Sty in the mind's eye: A meta-analytic investigation of the nomological network and internal consistency of the “Reading the mind in the eyes” test. *Assessment*, 29(5), 872–895. <https://doi.org/10.1177/1073191121996469>
- Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.
- Kynast, J., & Schroeter, M. L. (2018). Sex, age, and emotional valence: Revealing possible biases in the “Reading the mind in the eyes” task. *Frontiers in Psychology*, 9, 570. <https://doi.org/10.3389/fpsyg.2018.00570>
- Lee, H. R., Nam, G., & Hur, J. W. (2020). Development and validation of the Korean version of the Reading the mind in the eyes test. *PLoS One*, 15(8), Article e0238309. <https://doi.org/10.1371/journal.pone.0238309>
- Li, T.-S., Liu, C.-M., Liu, C.-C., Hsieh, M. H., Lin, Y.-T., Wang, E.-N., ... Chou, T.-L. (2020). Social cognition in schizophrenia: A network-based approach to a Taiwanese version of the Reading the mind in the eyes test. *Journal of the Formosan Medical Association*, 119(1), 439–448. <https://doi.org/10.1016/j.jfma.2019.08.008>
- Lindeman, M., Svedholm-Häkkinen, A. M., & Lipsanen, J. (2015). Ontological confusions but not mentalizing abilities predict religious belief, paranormal belief, and belief in supernatural purpose. *Cognition*, 134, 63–76. <https://doi.org/10.1016/j.cognition.2014.09.008>
- Macchia, A., Zebhauser, P. T., Salcedo, S., Burum, B., Gold, E., Alonso-Alonso, M., ... Brem, A.-K. (2021). Divergent effects of oxytocin on “mind-reading” in healthy males. *Cognitive, Affective, & Behavioral Neuroscience*. <https://doi.org/10.3758/s13415-021-00936-3>
- Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, 15, 51–69. <https://doi.org/10.1080/15366367.2017.1348108>
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 52(6), 2287–2305. <https://doi.org/10.3758/s13428-020-01398-0>
- McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, 28(1), 61–88. <https://doi.org/10.1037/met0000425>
- Megias-Robles, A., Gutierrez-Cobo, M. J., Cabello, R., Gomez-Leal, R., Baron-Cohen, S., & Fernandez-Berrocal, P. (2020). The “Reading the mind in the Eyes” test and emotional intelligence. *Royal Society Open Science*, 7(9), Article 201305. <https://doi.org/10.1098/rsos.201305>
- Mehrabian, A. (1996). *Manual for the balanced emotional empathy scale (BEES)* (Available from the author).
- Meyer, J. K., & Morey, L. C. (2015). Borderline personality features and associated difficulty in emotion perception: An examination of accuracy and bias. *Personality and Mental Health*, 9(3), 227–240. <https://doi.org/10.1002/pmh.1299>
- Milton, D., Gurbuz, E., & López, B. (2022). The “double empathy problem”: Ten years on. *Autism: The International Journal of Research and Practice*, 26(8), 1901–1903. <https://doi.org/10.1177/13623613221129123>
- Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18(1), 143. <https://doi.org/10.1186/s12874-018-0611-x>
- National Institute of Mental Health. (2016). Behavioral Assessment Methods for RDoC Constructs. <https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/understanding-mental-states.shtml>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69(1), 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Nemeth, N., Peterfalvi, A., Czeh, B., Tenyi, T., & Simon, M. (2020). Examining the relationship between executive functions and mentalizing abilities of patients with borderline personality disorder. *Frontiers in Psychology*, 11, 1583. <https://doi.org/10.3389/fpsyg.2020.01583>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., ... Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73(1), 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the mind in the eyes test. *Journal of Abnormal Psychology*, 125(6), 818–823. <https://doi.org/10.1037/abn0000182>
- Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brennehan, M. W., & Roberts, R. D. (2015). A psychometric analysis of the Reading the mind in the eyes test: Toward a brief form for research and applied settings. *Frontiers in Psychology*, 6, 1503. <https://doi.org/10.3389/fpsyg.2015.01503>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349. <https://doi.org/10.1126/science.aac4716>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (Online)*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Pavlova, M. A., & Sokolov, A. A. (2022). Reading language of the eyes. *Neuroscience and Biobehavioral Reviews*, 140, 104755. <https://doi.org/10.1016/j.neubiorev.2022.104755>
- Peñuelas-Calvo, I., Sareen, A., Sevilla-Lliewellyn-Jones, J., & Fernández-Berrocal, P. (2019). The “Reading the mind in the eyes” test in autism-spectrum disorders comparison with healthy controls: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders*, 49(3), 1048–1061. <https://doi.org/10.1007/s10803-018-3814-4>
- Peterson, E., & Miller, S. F. (2012). The eyes test as a measure of individual differences: How much of the variance reflects verbal IQ? *Frontiers in Psychology*, 3, 220. <https://doi.org/10.3389/fpsyg.2012.00220>
- Petrocchi, S., Filippini, C., Antonietti, C., Levante, A., & Lecciso, F. (2021). Theory of mind as a mediator between emotional trust beliefs and interpersonal communication competence in a group of young adults. *Psychological Reports*, 124(2), 555–576. <https://doi.org/10.1177/0033294120913489>
- Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2018). Social cognition psychometric evaluation: Results of the final validation study. *Schizophrenia Bulletin*, 44(4), 737–748. <https://doi.org/10.1093/schbul/sbx117>
- Pinkham, A. E., Penn, D. L., Green, M. F., Buck, B., Healey, K., & Harvey, P. D. (2014). The social cognition psychometric evaluation study: Results of the expert survey and RAND panel. *Schizophrenia Bulletin*, 40(4), 813–823. <https://doi.org/10.1093/schbul/sbt081>
- Pinkham, A. E., Penn, D. L., Green, M. F., & Harvey, P. D. (2016). Social cognition psychometric evaluation: Results of the initial psychometric study. *Schizophrenia Bulletin*, 42(2), 494–504. <https://doi.org/10.1093/schbul/sbv056>

- Platak, S. M., Keenan, J. P., Gallup, G. G., & Mohamed, F. B. (2004). Where am I? The neurological correlates of self and other. *Brain Research. Cognitive Brain Research*, 19(2), 114–122. <https://doi.org/10.1016/j.cogbrainres.2003.11.014>
- Preti, A., Vellante, M., & Petretto, D. R. (2017). The psychometric properties of the “Reading the mind in the eyes” test: An item response theory (IRT) analysis. *Cognitive Neuropsychiatry*, 22(3), 233–253. <https://doi.org/10.1080/13546805.2017.1300091>
- Prevost, M., Carrier, M. E., Chowne, G., Zolkowitz, P., Joseph, L., & Gold, I. (2014). The Reading the mind in the eyes test: Validation of a French version and exploration of cultural variations in a multi-ethnic city. *Cognitive Neuropsychiatry*, 19(3), 189–204. <https://doi.org/10.1080/13546805.2013.823859>
- Quesque, F., & Rossetti, Y. (2020). What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science*, 15(2), 384–396. <https://doi.org/10.1177/1745691619896607>
- Raffo De Ferrari, A., Lagravinese, G., Pelosin, E., Pardini, M., Serrati, C., Abbruzzese, G., & Avanzino, L. (2015). Freezing of gait and affective theory of mind in Parkinson disease. *Parkinsonism & Related Disorders*, 21(5), 509–513. <https://doi.org/10.1016/j.parkreldis.2015.02.023>
- Raimo, S., Trojano, L., Pappacena, S., Alaia, R., Spitaleri, D., Grossi, D., & Santangelo, G. (2017). Neuropsychological correlates of theory of mind deficits in patients with multiple sclerosis. *Neuropsychology*, 31(7), 811–821. <https://doi.org/10.1037/neu0000372>
- Redondo, I., & Herrero-Fernández, D. (2018). Validation of the Reading the mind in the eyes test in a healthy Spanish sample and women with anorexia nervosa. *Cognitive Neuropsychiatry*, 23(4), 201–217. <https://doi.org/10.1080/13546805.2018.1461618>
- Riem, M. M. E., Bakermans-Kranenburg, M. J., Voorthuis, A., & van Ijzendoorn, M. H. (2014). Oxytocin effects on mind-reading are moderated by experiences of maternal love withdrawal: An fMRI study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 51, 105–112. <https://doi.org/10.1016/j.pnpbp.2014.01.014>
- Russell, T. A., Schmidt, U., Doherty, L., Young, V., & Tchanturia, K. (2009). Aspects of social cognition in anorexia nervosa: Affective and cognitive theory of mind. *Psychiatry Research*, 168(3), 181–185. <https://doi.org/10.1016/j.psychres.2008.10.028>
- Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement*, 72(6), 910–932. <https://doi.org/10.1177/0013164412452564>
- Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology*, 5(1). <https://doi.org/10.1525/collabra.24>
- Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2), 65–72. <https://doi.org/10.1016/j.tics.2014.11.007>
- Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology*, 5. <https://doi.org/10.15626/MP.2019.1645>
- Schurz, M., Radua, J., Tholen, M., Maliske, L., Margulies, D., Mars, R., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, 147(3), 293–327. <https://doi.org/10.1037/bul0000303>
- Scott, L. N., Levy, K. N., Adams, R. B., Jr., & Stevenson, M. T. (2011). Mental state decoding abilities in young adults with borderline personality disorder traits. *Personality Disorders, Theory, Research, and Treatment*, 2(2), 98–112. <https://doi.org/10.1037/a0020011>
- Shaw, M., Cloos, L. J. R., Luong, R., Elbaz, S., & Flake, J. K. (2020). Measurement practices in large-scale replications: Insights from many labs 2. *Canadian Psychology/Psychologie Canadienne*, 61(4), 289–298. <https://doi.org/10.1037/cap0000220>
- Silverman, C. (2022). How to read “Reading the Mind in the Eyes.”. *Notes and Records of the Royal Society of London*, 76(4), 683–697. <https://doi.org/10.1098/rsnr.2021.0058>
- Slaney, K. (2017). *Validating psychological constructs historical, philosophical, and practical dimensions*. Palgrave Macmillan.
- Topić, M. K., & Kovačević, M. P. (2019). Croatian adaptation of the revised Reading the mind in the eyes test (RMET). *Psihologijske Teme*, 28(2), 377–395. <https://doi.org/10.31820/pt.28.2.8>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467–473. <https://doi.org/10.7326/M18-0850>
- Van den Stock, J., Stam, D., De Winter, F., Mantini, D., Szmeccsanyi, B., Van Laere, K., Vandenberghe, R., & Vandenbulcke, M. (2017). Moral processing deficit in behavioral variant frontotemporal dementia is associated with facial emotion recognition and brain changes in default mode and salience network areas. *Brain and Behavior: A Cognitive Neuroscience Perspective*, 7(12). <https://doi.org/10.1002/brb3.843>. e00843–n/a.
- Van Staden, J. G., & Callaghan, C. W. (2021). An evaluation of the reading the mind in the eyes test's psychometric properties and scores in South Africa-cultural implications. *Psychological Research*, 86(7), 2289–2300. <https://doi.org/10.1007/s00426-021-01539-w>
- Vassar, M., & Hale, W. (2009). Reliability reporting across studies using the buss Durkee hostility inventory. *Journal of Interpersonal Violence*, 24(1), 20–37. <https://doi.org/10.1177/0886260508314931>
- Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, 31(2), 162–168. <https://doi.org/10.1177/09637214211067779>
- Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., & Preti, A. (2013). The “Reading the mind in the eyes” test: Systematic review of psychometric properties and a validation study in Italy. *Cognitive Neuropsychiatry*, 18(4), 326–354. <https://doi.org/10.1080/13546805.2012.721728>
- Veritas Health Innovation. (2022). *Covidence systematic review software*.
- Warrier, V., Grasby, K. L., Uzevovsky, F., Toro, R., Smith, P., Chakrabarti, B., ... Baron-Cohen, S. (2018). Genome-wide meta-analysis of cognitive empathy: Heritability, and correlates with sex, neuropsychiatric conditions and cognition. *Molecular Psychiatry*, 23(6). <https://doi.org/10.1038/mp.2017.122>
- Weinreb, S., Li, F., & Kurtz, M. M. (2022). A meta-analysis of social cognitive deficits in schizophrenia: Does world region matter? *Schizophrenia Research*, 243, 206–213. <https://doi.org/10.1016/j.schres.2022.04.002>
- Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, 55(2), 788–806. <https://doi.org/10.3758/s13428-022-01849-w>
- Wolf, M. G. (2023). The problem with over-relying on quantitative evidence of validity. *PsyArXiv*. <https://doi.org/10.31234/osf.io/v4nb2>
- Yildirim, E. A., Kasar, M., Guduk, M., Ates, E., Kucukparlak, I., & Ozalmete, E. O. (2011). Investigation of the reliability of the “Reading the mind in the eyes test” in a Turkish population. *Türk Psikiyatri Dergisi*, 22(3), 177–186. <https://doi.org/10.5080/u6500>
- Zamoscik, S., S. N. L., Bravo, R., Ugartemendia, L., Plieger, T., Rodriguez, A. B., ... Kirsch, P. (2021). Tryptophan-enriched diet or 5-hydroxytryptophan supplementation given in a randomized controlled trial impacts social cognition on a neural and behavioral level. *Scientific Reports*, 11(1), 21637. <https://doi.org/10.1038/s41598-021-01164-y>
- Zhong, J., Zhu, H., Yin, D., Ning, Y., Zheng, S., Zhang, Y., & Jia, H. (2021). Paliperidone compared with haloperidol on the theory of mind tasks in schizophrenia: A pilot trial. *Neuropsychiatric Disease and Treatment*, 17, 3683–3691. <https://doi.org/10.2147/NDT.S335597>