The "Reading the Mind in the Eyes" Test shows poor psychometric properties in a large, demographically representative US sample

Wendy C. Higgins[a], Robert M. Ross[b], Robyn Langdon[a], and Vince Polito[a]

[a] Macquarie University, Department of Cognitive Science, NSW, 2109, Australia
[b] Macquarie University, Department of Psychology, NSW, 2109, Australia

This manuscript has now been published in *Assessment.* Please consult the published version:

Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2022). 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/10.1177/10731911221124342

Author Note

Wendy C. Higgins https://orcid.org/0000-0003-1357-8330

Robyn Langdon

Vince Polito https://orcid.org/0000-0003-3242-9074

Robert M. Ross https://orcid.org/0000-0001-8711-1675

We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Wendy C Higgins. Email: wendy.higgins@unimelb.edu.au

Declarations of interest: none

Acknowledgments

Wendy Higgins was supported by an Australian Government Research Training Program (RTP) Scholarship, a Macquarie University Research Excellence Scholarship, and a Macquarie University Cognitive Science Postgraduate Grant. And Robert Ross was supported by the Australian Research Council (DP180102384) and a Macquarie University Research Fellowship.

Abstract

The Reading the Mind in the Eyes test (RMET) is a widely used measure of theory of mind (ToM). Despite its popularity, there are questions regarding the RMET's psychometric properties. In the current study, we examined the RMET in a representative US sample of 1,181 adults. Key analyses included conducting an exploratory factor analysis on the full sample and examining whether there is a different factor structure in individuals with high versus low scores on the 28-item autism spectrum quotient (AQ-28). We identified overlapping, but distinct, three-factor models for the full sample and the two subgroups. In all cases, each of the three models showed inadequate model fit. We also found other limitations of the RMET, including that nearly a quarter of the RMET items did not meet the criteria for inclusion in the RMET that were established in the original validation study. Due to the RMET's weak psychometric properties and the uncertain validity of individual items, as indicated by our study and previous studies, we conclude that significant caution is warranted when using the RMET as a measure of ToM.

Keywords: Reading the Mind in the Eyes Test, construct validity, theory of mind, factor analysis, autism quotient

1. Introduction

Broadly defined, theory of mind (ToM) is the ability to impute mental states to oneself and others (Premack & Woodruff, 1978). Baron-Cohen (1995) proposed that a deficit in ToM ability is an underlying cause of autism. Noting an absence of ToM tests suitable for use with adults, Baron-Cohen and colleagues developed a new measure, the Reading the Mind in the Eyes test (RMET, Baron-Cohen et al., 1997, Baron-Cohen, Wheelwright, Hill et al., 2001), to identify subtle ToM deficits in adults. Specifically, the RMET was designed to measure the ability to use information from people's eyes to infer mental states as an indication of ToM ability.

1.1 The RMET

The RMET comprises 36-items, with each item consisting of a black and white photograph of a person's eyes and four mental state descriptors which the authors selected to be of a similar valence (e.g., terrified, upset, arrogant, annoyed).[1] Participants are instructed to select "the word that best describes what the person in the picture is thinking or feeling" (see Figure S1). The photographs were collected from magazines and the mental state descriptors were selected by the authors of the test. The individual test items were validated in a sample of 225 participants in Britain, consisting of 122 members of the general public and 103 undergraduate students from Cambridge University. The validation criteria for each RMET item were that (1) at least 50% of participants chose the target response and (2) no more than 25% selected the same incorrect response. The RMET is made up of the 36 out of 40 items that met these validity criteria make up

Notes

[1] The original version of the task contained 25 items and paired two mental state descriptors with opposite meanings to each image (Baron-Cohen et al., 1997), but the test was revised because the original version was found to be too easy (Baron-Cohen, Wheelwright, Hill et al., 2001). We will not discuss the original 25-item version further and for clarity of exposition will refer to the 36-item version as "the" RMET.

the RMET. It is widely used in empirical research and the paper that introduced it is extensively cited, having 3,348 citations on Web of Science as of May, 23 2022.

Baron-Cohen, Wheelwright, Hill et al. (2001) did not evaluate the RMET's test-retest reliability or internal consistency, and these statistics are rarely reported in subsequent studies using the RMET. When reported, the test-retest reliability of the RMET is generally acceptable (e.g., Fernández-Abascal et al., 2013; Hallerbäck et al. 2009; Khorashad et al., 2015; Prevost et al., 2014). Conversely, there is considerable variation in reported levels of internal consistency based on Cronbach's alpha (α), which is the most frequently reported measure of internal consistency for the RMET (Kittel et al., 2021). While some studies have reported moderate (e.g., Kuczynski et al., 2020; Soker-Elimaliah et al., 2020) to high levels of internal consistency (e.g., Israelashvili et al. 2020; Ozturk et al., 2020), other studies have reported low levels of internal consistency (e.g., Giordano et al., 2019; Khorashad et al., 2015; Koo et al., 2021; Meyer & Shean, 2006; Vellante et al., 2013). A recent meta-analysis of the psychometric properties of the RMET reported an acceptable α value of .73 (95% CI [.65, .79]) based on 21 effect sizes (Kittel et al., 2021). However, as the authors note, longer psychometric instruments such as the RMET can inflate α values.

The variation in levels of internal consistency reported for the RMET might relate to the use of α as a measure of internal consistency because α relies on the assumption that a test is unidimensional (Goodboy & Martin, 2020; Olderbak et al., 2021). While the RMET was designed to evaluate a single ability (attributing mental states based on images of eyes) as in indicator of ToM capacity, Baron-Cohen, Wheelwright, Hill et al. (2001) did not evaluate the test's factor structure, and subsequent research has yet to clearly establish its factor structure. To our knowledge, Olderbak et al., (2015) are the only researchers to conduct an exploratory factor analysis (EFA) on the 36-item English version of the RMET. They identified a five-factor model. However, they rejected the five-factor model because nine items (i.e., 25% of the items) failed to load on to any of the factors, the factors were weakly related to each other, and none of the factors related to any theoretically-motivated subscales previously proposed in the literature. Using confirmatory factor analysis (CFA),

they evaluated two theoretically-driven models, a single-factor model and a three-factor model proposed by Harkness et al. (2005), in which test items were divided into positive, negative, and neutral factors based on the valence of each item. However, key measures of model fit provided inconsistent results and the factor loadings were low, so the authors ultimately rejected both the single-factor and the three-factor valence models.

We are only aware of three other studies that conducted factor analyses on the English language RMET. Black (2019) also conducted CFA on a single-factor model. While the author concluded that the single-factor model had acceptable model fit, similar to the results reported by Olderbak et al. (2015), the reported model fit statistics were inconsistent. Individual factor loadings were not reported. It is difficult to reconcile the findings of Olderbak et al. (2015) and Black (2019) because, despite similar results from a CFA on a single-factor model for the RMET, these two studies came to opposite conclusions about the acceptability of the model. Nonetheless, Kline (2016, p. 264) emphasises the importance of evaluating model fit against multiple fit indices, “Because a single statistic reflects only a particular aspect of fit, a favorable value of that statistic does not by itself indicate acceptable fit.” This supports the rejection of the single factor model due to the inconsistent results across key model fit measures. Two other studies (Benau et al., 2020; Sherman et al., 2020) briefly report the results of a CFA for a single factor model as evidence for acceptable internal consistency of the RMET. While the authors of both studies stated that they found good model fit, the fit statistics that they provided were not comprehensive and did not include the fit statistics that showed poor model fit in the analyses conducted by Olderbak et al. (2015) and Black (2019). This lack of detail limits our ability to evaluate their claim of good model fit. Overall, the lack of clear evidence in support of a single factor model for the RMET is of concern because, as noted above, the most frequently reported measure of internal reliability for the RMET is α, which assumes unidimensionality (Goodboy & Martin, 2020; Olderbak et al., 2021). If the RMET is not unidimensional, then α is not appropriate as a measure of its internal consistency.

In line with the proposition that ToM deficits are an important underlying cause of autism (Baron-Cohen, 1995), Baron-Cohen, Wheelwright, Hill et al. (2001) validated the RMET as a measure of ToM based on differences in RMET performance between autistic and non-autistic participants and correlations between RMET scores and scores on a measure of autistic traits, the Autism Spectrum Quotient (AQ; Baron-Cohen, Wheelwright, Skinner et al., 2001). The authors found that autistic individuals[2] scored significantly lower on the RMET than non-autistic individuals and that RMET scores negatively correlated with AQ scores in both autistic and non-autistic participants. While these findings offered initial support for the validity of the RMET as a measure of ToM based on the proposed relationship between autism and ToM, establishing the validity of the RMET as a measure of ToM through the differential performance of autistic individuals and individuals with high levels of autistic traits is potentially problematic for at least three reasons.

First, it relies on the assumption that autistic individuals (and individuals with higher levels of autistic traits) have a deficit in ToM ability[3], which is not without contention. For example, Deschrijver and Palmer (2020) have proposed that atypical performance in ToM tasks in autistic individuals might relate to differences in how they process conflict between their own mental states and the mental states of others, rather than a problem with imputing mental states per se.

---

[2] A note on language: In this paper we use identify first terminology to refer to autistic individuals. While there are mixed language preferences within the autism community, it has been argued that identify first language tends to be more preferred by the autism community than person first language (Botha et al., 2021).

[3] Notably, lower RMET scores have also been associated with clinical disorders in which multiple cognitive capacities that are impacted in addition to ToM, including schizophrenia (Pinkham et al., 2016), anorexia nervosa (Russell et al., 2009), traumatic brain injury (Fazaeli et al., 2018; Muller et al., 2010), euthymia and bipolar disorder (Bora et al., 2016). This illustrates that although poor performance of autistic individuals on the RMET might be consistent with the RMET being a measure of ToM, without additional types of converging evidence, the sensitivity of the RMET to clinical conditions alone is not sufficient to establish the RMET as a valid measure of ToM.

Moreover, Gernsbacher and Yergeau (2019) recently challenged the body of research purportedly establishing the existence of ToM deficits in autistic individuals.

Second, it relies on the assumption that autistic individuals' lower scores on the RMET are caused by a deficit in ToM ability. However, the RMET might tap into other abilities or dispositions that result in autistic individuals having low scores. For example, there is evidence that the RMET measures emotion recognition abilities, and alexithymia (a condition in which an individual has difficulty recognising and describing their own and others' emotions) frequently co-occurs with autism (Bird & Cook 2013). Importantly, Oakley et al. (2016) recently found that alexithymia is more predictive of performance on the RMET than an autism diagnosis or autistic traits measured using the AQ. Consequently, they concluded that differences in autistic individuals' RMET performance are better explained by atypical emotion recognition related to co-occurring alexithymia than by differences in ToM ability. Additionally, a recent meta-analysis revealed that RMET performance is more strongly associated with performance in measures of emotion perception than measures of ToM (Kittel et al., 2021). There is also evidence that RMET performance correlates with vocabulary (Kittel et al. 2021; Olderbak et al., 2015), suggesting that it indexes crystalised intelligence.

Third, the RMET might not be an appropriate measure of ToM ability for autistic individuals due to the nature of the stimuli. There is evidence that autistic individuals can find looking directly at other people's eyes uncomfortable or even overwhelming (Hadjikhani et al., 2017; Stuart et al., 2022; Trevisan et al., 2017). Thus, some autistic individuals may find the stimuli used in the RMET aversive, predisposing them toward poorer performance on the task, regardless of their underlying ToM ability. Autistic traits, as indexed by the AQ, have been shown to be normally distributed within the population (Hurst et al., 2007; Ruzich et al., 2015) and there is evidence that AQ scores correlate negatively with RMET scores in the general population (Gökçen et al., 2016; Kallitsounaki & Williams, 2020). Thus, the influence of the stimuli on performance might also extend to individuals who exhibit high levels of autistic traits. This possibility is supported by recent research showing that

members of the general population study who scored higher on the AQ spend more time looking at the bottom half of a face than those with lower AQ scores (Wegner-Clemens et al., 2020).

Despite the limited evidence of the validity of the RMET, it is regularly claimed that the RMET is a "well-validated measure" without providing supporting evidence (see Table S1 for a list of quotations). This illustrates a form of research bias in the psychological sciences that Flake and Fried (2020) have recently dubbed a "measurement schmeasurement" attitude that is characterised by the use of psychometric instruments without sufficient evidence for their validity. A re-evaluation of the validity of the RMET is critical because, as Flake and Fried note, if a test is not a valid measure of its target construct, then the inferences drawn from the study are invalid and "[n]either rigorous research design, nor advanced statistics, nor large samples can correct such false inferences" (Flake & Fried, 2020, p. 456). Thus, if it turns out that the RMET is not a valid measure of ToM, then the consequences for the sprawling body of literature using this measure are profound.

As a step toward addressing the limited evidence of the validity of the RMET, the current study evaluated the RMET's factor structure. We also examined the possibility that individuals with higher versus lower levels of autistic traits might exhibit different factor structures due to aspects of the test that are unrelated to ToM ability. For example, if individuals high in autistic traits find viewing the RMET stimuli to be uncomfortable and/or rely on compensatory strategies for identifying the correct response (Baron-Cohen, Wheelwright, Hill et al., 2001; Golan et al., 2015; Livingston et al., 2020) this might help explain inconsistencies in earlier factor analyses.

1.2 Aims and Hypotheses

The present study was pre-registered, with six primary aims. First, to evaluate the factor structure of the RMET in a large demographically representative US sample. Second, to test whether there are different factor structures in individuals with higher versus lower levels of autistic traits (as measured using an abbreviated, 28-item version of the AQ; AQ-28; Hoekstra et al., 2011). Third, to test whether individuals with higher levels of autistic traits who score lower on the RMET do so as a result of co-occurring alexithymic traits (as measured by the Toronto Alexithymia Scale, TAS, Bagby

et al., 1994). Fourth, to evaluate the convergent validity of the RMET with another measure of ToM, the Imposing Memory Test (IMT, Launay et al., 2015). Fifth, to examine relationships between performance on the RMET, autistic traits, and the level of comfort participants reported when viewing the eye stimuli (using an ad hoc single-item measure of level of comfort measure). Sixth, to examine the factor structures of the TAS and the AQ-28.

We report full results for our first four aims in the main paper. We provide a brief summary of our results testing our fifth aim in the main paper, while detailed results are available in the supplementary materials (Part 2, section 2). All information related to the sixth aim are available on the project's Open Science Framework (OSF) page[4].

2. Method

We report how we determined our sample size, all data exclusion criteria, all manipulations, and all measures in the study. The pre-registration, data, R scripts, and supplementary materials are available on the project's OSF page[4].

2.1 Participants

The sample size for this study was determined based on Goretko et al.'s (2019) recommended minimum of 400 participants for exploratory factor analysis (EFA) when the number of factors is unknown. Because we planned to conduct separate EFAs on two subsets of our data consisting of (a) participants with the lowest third of AQ-28 scores and (b) participants with the greatest third of AQ-28 scores, our target sample size was 1,200 participants.

Participants were recruited using Lucid Theorem, an online recruitment platform that uses quota sampling to provide a sample that matches the US national distribution in terms of age, gender, ethnicity, and geographic region (Coppock & McClellan, 2019). Participants were compensated directly with cash, gift cards, or loyalty reward points by Lucid's partner companies according to the terms of their agreements with these partner companies. 2,678 participants were

[4] Project's OSF page: https://osf.io/8jtn9/

recruited to the study. We had three pre-registered exclusion criteria. First, for analyses involving gender, only participants who selected "male" or "female" gender categories were analysed (to ensure that the categories were large enough for statistical analysis). No participants were excluded based on this criterion because all participants identified as male or female in the demographic information they provided to Lucid Theorem. Second, 863 participants failed an attention check question that was presented near the end of the survey in which they were asked to show that they have read the questions by moving a slider to "0"[5]. These participants were immediately sent to a debriefing page and were excluded from all analyses. Third, 39 participants who did not finish the IMT were excluded from the analyses involving this measure. Additionally, examination of the data revealed that 41 participants provided straight line responses to the AQ-28, TAS, and/or IMT, which means that they selected the same response across all items for at least one of these measures. Because approximately half of the items on these measures are reversed scored, it is extremely unlikely that these are sincere responses, so these participants were excluded too. This was not a pre-registered exclusion criterion. Results with these participants retained, which are very similar to the results reported in the main paper, are included in Part 3 of the supplementary materials.

The final sample included 1,181 (652 female) participants with ages ranging from 18 to 88 (M = 47.7, SD = 17.0). The sample was representative of the US population in terms of levels of

[5] While we have excluded a large number of participants due to inattentiveness, it should be noted that research suggests that (i) participants recruited online tend to be more attentive than participants in the laboratory (Hauser & Schwarz, 2016; Ramsey et al., 2016) and (ii) excluding inattentive participants increases statistical power (Oppenheimer, Meyvis, & Davidenko, 2009) and attenuates spurious associations between behavioural and self-report measures (Sulik et al., 2021; Zorowitz et al., 2021). Given that it is very rare for laboratory-based studies using the RMET to include any attention checks, it is possible that laboratory-based RMET studies have included substantial numbers of unidentified inattentive participants, which could result in undetected spurious findings in the literature.

educational attainment (high school 23.5% [USA 28.3%][6], some university study 20.1% [17.7%], two-year degree 8.4% [9.8%], four-year degree 25.9%, [21.2%], postgraduate studies 13.2% [10.8%]) and race (White 76.6% [76.3%], Black 9.3% [13.4%], Asian 2.7% [5.9%], other 9.3%, and prefer not to answer 3.7%).

2.2 Measures

2.2.1 Reading the Mind in the Eyes Test (Revised Version)

The original revised RMET was a pen and paper test with one test item per page and the mental state terms printed around the four corners of the image (e.g., flustered, convinced, desired, joking). To avoid any response bias due to the placement of response options, we presented the four mental state descriptors below the image with the order of response options randomised across participants (Figure S1). The original version of the test includes a glossary of terms to ensure that participants know the meaning of all the mental state descriptors. Accordingly, for each RMET question, we included an "extra help" section that participants could click to see the definitions of the mental state descriptors for that test item. Items were scored with 1 point for a correct response and 0 points for an incorrect response, with the total possible score ranging from 0 to 36. Higher scores purportedly indicate higher levels of ToM ability.

2.2.2 Autism Spectrum Quotient Brief Version (AQ-28)

The AQ-28 (Hoekstra et al., 2011) is a 28-item reduced version of Baron-Cohen, Wheelwright, Skinner et al.'s (2001) 50-item Autism Spectrum Quotient questionnaire. In this self-report measure, participants rate their level of agreement with statements about themselves. Each item is scored from 1 to 4, (1 = "definitely agree''; 2 = "slightly agree"; 3 = "slightly disagree" and 4 = "definitely disagree"), with the total possible scores ranging from 28 to 112. Higher scores indicate more autistic traits. Hoekstra et al. (2011) found that the AQ-28 has acceptable internal consistency

[6] The numbers in square brackets are the actual US statistics according to the United States Census Bureau (2019a, 2019b).

(α = .78) and correlates highly with the original 50-item scale (*r* = .93). Hoekstra et al. (2011) identified a five-factor structure (social skills, routine, switching, imagination, and numbers and patterns) and a two-factor higher-order factor structure (numbers and patterns and social behaviour, which incorporates the other four first-order factors).

2.2.3. Toronto Alexithymia Scale (TAS)

The TAS (Bagby et al., 1994) is a self-report measure of alexithymia that comprises 20 statements. The TAS has three subscales, difficulty identifying feelings (identify), difficulty describing feelings (describe), and externally oriented thinking (external). Bagby et al. (1994) reported acceptable levels of test re-test reliability ($r = .77$, $p < .01$) and internal consistency (with the exception of the external subscale): α for full test = .81, identify = .78, describe = .75, external = .66. Each item was scored from 1 to 5, with total possible scores ranging from 20 to 100 and higher scores indicating higher levels of alexithymic traits.

2.2.4. The Imposing Memory Task (IMT)

The IMT (originally created by Kinderman et al., 1998) is a story test similar to the more widely used Strange Stories task (Happé, 1994). For this study, we used a single story of approximately 200 words from Launay et al.'s (2015) version of the IMT (adapted from Stiller & Dunbar, 2007). Participants read the story and then answered true/false questions based on its content. Some of the true/false questions required mental state reasoning (e.g., "Carolyn thought that Hannah liked Emma's boyfriend Matt"), whereas others were memory control questions (e.g., "Carolyn told Hannah that Emma had been at training"). We used a subset of 16 of the 22 questions related to the story, comprising eight ToM questions and eight memory control questions. Questions scored 1 point for a correct response and 0 points for an incorrect response. Both memory and ToM scores were calculated with a range of 0-8 for each category.

We made two small amendments to the questions by replacing a pronoun with a character's name to reduce ambiguity and replacing the word "friend" with the word "colleague" in another question as it was not clear from the story that the two characters were friends.

2.3 Procedure

Participants completed the survey online. The median completion time was 18 minutes. After the consent page, the RMET, TAS, and AQ-28 were presented with their order randomised across participants. The question about comfort viewing eye stimuli was always presented immediately after the RMET. Following these tasks, the IMT was presented. Finally, participants were asked demographic questions, which included a measure of belief in God that will be analysed as part of a separate project that we will report elsewhere. Ethics approval for this study was granted by [Redacted] Human Research Ethics Committee (reference number: 52020625515320).

2.4 Analytic Approach

Model fit was assessed using seven metrics: χ^2, root mean square of residuals (RMSR), standardised root mean square of residuals (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and Bayesian information criterion (BIC). A non-significant χ^2 indicates good model fit, however, as sample sizes increase, χ^2 becomes significant independent of model fit (Bergh, 2015). We have reported χ^2 for all models, however, because our sample size was very large, it is not at all surprising that χ^2 was always significant. CFI and TLI are relative fit measures, which means that they compare model fit to a null model. Higher values indicate better model fit. In contrast, RMSR, SRMR, and RMSEA are absolute fit measures, and model fit is evaluated without comparison to a null model. Lower values indicate better model fit. Lower values also indicate better model fit for BIC, and the value can be used to select between competing models.

Model fit for EFA and SEM analyses were determined according to the following fit criteria: RMSR < .05, SRMR < .08, RMSEA < .08 acceptable fit, < .05 good fit, TLI ≥ .95, CFI ≥ .90, χ^2, $p > .05$, and lower values indicate better model fit for BIC (Hu & Bentler, 1999). However, it should be noted that these values are guidelines and were not designed as strict cutoff criteria (Kline, 2016; Marsh et al., 2004). In addition to evaluating models according to these metrics, we made decisions based on the conceptual applicability of the models.

Omega (ω) is recommended as a more appropriate indicator of internal consistency than α (Flora, 2020). In contrast to α, which assumes unidimensionality (Goodboy & Martin, 2020; Olderbak et al., 2021), ω provides information related to a measure's dimensionality as well as its internal consistency. However, ω requires knowledge of the factor structure of a measure (Flora, 2020). As noted above, previous research has not been able to identify a well-fitting factor model for the RMET. Because we also could not find a well-fitting factor model for the RMET, we report α based on the test's proposed single factor (Baron-Cohen, Wheelwright, Hill et al., 2001) and to allow for direct comparison with previous studies, which have predominantly reported α values for the RMET. In line with the findings of Black (2019) and Olderbak et al. (2015), an exploratory CFA on our data resulted in inconsistent fit statistics (see supplementary materials, Part 2, section 1). Following Flora's (2020) guidelines, we report omega hierarchical (ω_h) as a measure of internal consistency for the AQ-28 and the TAS because the measures' proposed multidimensional structures are supported by CFA model fit statistics. We also report α for these measures for comparison with previous studies. There are no strict cutoff values for these measures to indicate acceptable levels of internal consistency (Green & Yeng, 2015). Nonetheless, minimum values of ≥ .70 are often cited for α (Christmann & Van Aelst, 2006) and researchers have also used the value of ω ≥ .70 as indicative of acceptable reliability (Bado et al., 2018). We also report the mean inter-item tetrachoric correlation for the RMET, which indicates the level of agreement between test items. The recommended range is .15-.50 (Clark & Watson, 1995).

3. Results

3.1 Descriptive Statistics

Tables S2 and S3 contain the descriptive statistics for the outcome variables and the correlation matrix for all key variables. RMET scores had a slight left skew (Figure S2), but the level of skew was within the acceptable range to assume a normal distribution (skew = -0.73; kurtosis = 0.35; Hair et al., 2014; Hancock et al., 2018). Scores ranged from 5 (14% correct) to 34 (94% correct). The mean score (*M* = 23.49, 65% correct, *SD* = 5.51) was comparable, but lower than in the general

population reported by Baron-Cohen, Wheelwright, Hill et al. (2001, *M* = 26.2 *SD* = 3.6). There were eight test items that failed Baron-Cohen, Wheelwright, Hill et al.'s (2001) criteria for validating test items: specifically, greater than 25% of participants selected the same foil for items 6, 10, 17, 23, 25, 28, 34, and 35; and less than half of participants selected the correct response for items 23 and 25 (see Table S4). Reliability was acceptable according to α at .75. We also found a weak positive correlation between reported levels of comfort viewing the eye stimuli and RMET performance (r = .19, $p < .001$).

The mean inter-item tetrachoric correlation for the RMET was .13 (range from -.12 to .36, see Table S5 for the full correlation matrix) which, consistent with previously reported values (.10, Black, 2019; .08, Olderbak et al., 2015), falls below the recommended range of .15-.50 (Clark & Watson, 1995). A subset of items were also negatively correlated with each other. This indicated low levels of agreement between items. Also, in line with Olderbak et al. (2015), correlations between items with same target were low ("fantasizing" r = .21, "cautious" r = .10, "preoccupied" r = .32, "interested" r = .15).

AQ-28 scores were normally distributed (Figure S3). Scores ranged from 39 to 102 (*M* = 66.0, *SD* = 9.5). Internal consistency was within the acceptable range for the full scale (ω_h = .80; α = .77) and the *social skills* (ω_h = .71; α = .82) and the *numbers and patterns* subscales (ω_h = .72; α = .72). However, the reliability of the other three subscales were below the recommended range (*routine* [ω_h = .46; α = .57], *switching* [ω_h = .57; α = .57], *imagination* [ω_h = .66; α = .69]).

TAS scores were normally distributed. Scores ranged from 22 to 83 (*M* = 49.4, *SD* = 12.3). Internal consistency was within an acceptable range for the full scale (ω_h = 0.87, α = .85) and the TAS *describe* and TAS *identify* subscales (ω_h = .78, α = .77; ω_h = .85, α = .86 respectively). However, consistent with Bagby et al. (1994, 2014), the internal consistency of the TAS *external* subscale was low (ω_h = .47, α = .54).

3.2 EFA of Overall RMET Factor Structure

We used EFA to evaluate the factor structure of the RMET. An item was considered to load onto a factor if the rotated factor weighting was ≥ 0.3 (Hair et al., 2014). Model fit was evaluated against the criteria outlined in section 2.4.

For the full sample, we ran parallel analysis to determine the number of factors to retain, using the *psych* package (version 2.0.9, Revelle, 2020) in R (version 4.0.1, R Core Team, 2020). Because the items are dichotomous, we used a tetrachoric correlation matrix, weighted least squares factoring method with an oblique rotation (geominQ) and 50 iterations (Susana et al., 2017). Parallel analysis suggested 12 factors, and model fit measures for this solution approached good fit levels (CFI = .889, TLI = .732, RMSEA = .051, 95% CI [.045-.055], RMSR = .02, BIC = -782, χ^2 = 1086, p < .001). However, in this model, 11 items failed to load on to any factor, seven factors consisted of only a single item, and the other five factors lacked any obvious conceptual explanatory power.

Because there was a poor conceptual fit for the model retaining the number of factors indicated by parallel analysis, we conducted an exploratory analysis using Cattell's scree plot[7] (Figure S3). This approach indicated a three-factor solution. This model had acceptable fit when evaluated by RMSEA (.059, 95% CI [.057-.061]) and good model fit according to RMSR (.05) and χ^2 (2693, *p* <.001.) However, model fit was poor according to global fit indices (CFI = .706, TLI = .647, BIC = -1021). Similar to the model with more factors retained we found that nine items did not load onto any of the three factors, three items had cross loading on two factors, and the maximum factor

[7] Another method of determining how many factors to retain is to retain only factors with eigenvalues ≥ 1. That method indicated a two-factor solution. The three-factor model based on the scree plot was preferred because it had better model fit indices and better conceptual explanatory power. However, both the two and three-factor models suffered from the same limitations of overall poor model fit, low factor loadings, and a high number of items failing to load on to any factor, which indicated that ultimately, both models should be rejected.

loading was only 0.550 (see Table 1). The cumulative variance explained was also low (.20), indicating that a significant amount of variance is not explained by this model.

Despite the poor model fit and high number of items failing to load onto any factor, the three-factor solution did have conceptual explanatory power, with one factor relating to internally oriented attention and thinking (e.g., pensive, preoccupied), one factor relating to negative emotions (e.g., hostile, despondent), and one factor relating to flirtation (e.g., flirtatious, fantasizing). The flirtatious factor overlaps with one of the five factors identified by Olderbak et al. (2015), with five items in common: “desire” (3), “flirtatious” (30), “fantasizing” (21, 6), “interested” (25). However, a number of items failed to load as would be expected. The target “reflective” (29) did not load on to the thoughtful factor. The targets “upset” (2), “worried” (5), and “accusing” (14) did not load on to the negative factor. Only one of the two RMET items with the target “interested” loaded on to the flirtatious factor.

3.3 EFA of RMET Factor Structure in Participants Low in Autistic Traits

To evaluate the possibility that individuals with lower levels of autistic traits show a different factor structure for the RMET, we conducted EFA on the RMET scores of the participants with the highest and lowest AQ-28 scores. While there are no strict guidelines for the minimum sample size for EFA, Goretzko et al. (2019) recommend a minimum of 400 participants when number of items per factor and the amount of variance the factors will account for are unknown. In line with this recommendation and because there is currently little information related to the factor structure of the RMET, we specified in our pre-registration that we would select the top and bottom third of participants based on AQ-28 scores to ensure adequate group sizes for robust factor analyses. Scores for participants in the bottom third ranged from 39 to 62 (*M* = 56.2, *SD* = 4.8). This subgroup consisted of 422 (233 female) participants with a mean age 49.2 (*SD* = 16.1). This subgroup was demographically similar to the full sample (White 78%, Black 11%, Asian 3%, other 6%, and prefer not to answer 2%). The mean RMET score was 24.2 (*SD* = 4.7). Internal consistency was below the

"acceptable" range for α at .68, and some items negatively correlated with the scale, suggesting that these items do not all represent a single factor.

Parallel analysis suggested 13 factors, however, the 13-factor model did not converge. In fact, no factor solution from 1 to 13 resulted in good model fit. Cattell's scree plot (Figure S3) indicated a three-factor solution, but all fit measures indicated poor model fit (CFI = .430, TLI = .312, RMSEA = .094, 95% CI [.090-.098], RMSR = .07, BIC = -704, χ^2 = 2469, p < .001), 13 items did not load onto any factor, and three items had cross-loadings (see Table 2). The three factors overlapped considerably with the results from the full sample, and conceptually matched the division into thoughtful, negative, and flirtatious factors. The maximum factor loading of .615 was higher than for the full sample. The cumulative variance explained was comparable to the full sample (.19).

Five items failed to load as expected on to the negative factor. The items "upset" (2), "worried" (5), and "uneasy" (7) did not load on to any factor, while "distrustful" (34), and "sceptical" (12), which loaded on to the negative factor in the full sample loaded on to the thoughtful factor in the low AQ-28 scores subgroup. Additionally, "fantasizing" (6) did not load on to the flirtatious factor.

3.4 EFA of RMET Factor Structure in Participants High in Autistic Traits

Scores for the group of participants with AQ-28 scores in the top third ranged from 70 to 102 (*M* = 76.1, *SD* = 5.9). This subgroup consisted of 409 (234 female) participants with a mean age of 45.7 (*SD* = 17.0). Demographics were similar to the full sample (White 74%, Black 10%, Asian 3%, other 5%, and prefer not to answer 5%). Mean RMET score for this subgroup was 23.5 (*SD* = 5.7). Internal consistency of the RMET was acceptable according to α at .78.

Parallel analysis suggested 14 factors; however, the 14-factor model did not converge. Cattell's scree plot (Figure S3) indicated a three-factor model, but all fit indices indicated poor model fit (CFI = .486, TLI = .379, RMSEA = .107, 95% CI [.104-.111], RMSR = .07, BIC = -154, χ^2 = 3003, p < .001), eight items did not load on to any factor, and four items had cross loadings on two factors (see Table 3). The high AQ-28 subgroup had the highest maximum factor loading (0.807), and the

factor structure for this subgroup was notably different from the full sample and the low AQ-28 subgroup. Factor 1 had 21 items loaded onto it. Factor 2 only had three items, and two of these items had cross loadings. Factor 3 only had four items, one of which had a cross loading on Factor 1. The cumulative variance explained was 0.25.

3.5. Evaluation of the construct validity of the RMET

To assess (1) whether alexithymia traits and/or autistic traits are associated with RMET performance and (2) whether the RMET performance correlates with the IMT, another measure of ToM, we used structural equation modelling (SEM) to evaluate relationships between performance on the RMET and performance on the TAS, AQ-28, and IMT, while controlling for gender (gender has been shown to be associated with performance on the RMET in individuals without an autism diagnosis, [Baron-Cohen et al., 2015]). In line with the pre-registration, 39 participants who did not complete the IMT were excluded from this analysis, leaving 1,142 (624 female) participants.

We tested a SEM model with paths to the RMET from gender, TAS subscale scores, the AQ-28 first-order subscale scores, IMT memory scores, and IMT ToM scores. The resulting model had 0 degrees of freedom, indicating a saturated model. This means that model fit could not be evaluated, and the SEM resulted in a multiple linear regression of RMET scores on the variables (see Table 4), which indicated that all three TAS subscales and the AQ-28 *imagination* and *social skills* subscales correlated with RMET scores, as did the IMT memory and ToM scores. AQ-28 total score was not significantly correlated with RMET scores (see Table S3).

4. Discussion

4.1 Factor Structure of the RMET

In this study we evaluated the factor structure of the RMET in a demographically representative US sample of 1,181 participants. We hypothesised that the RMET is a multidimensional measure of ToM ability and conducted EFA to identify the hypothesised multifactorial structure. Consistent with Olderbak et al. (2015), we failed to identify an appropriate factor structure for the RMET. The best statistical model fit was obtained by retaining 12 factors;

however, this model was not viable because over half of the factors only contained a single item and for those factors that did contain multiple items, there was no obvious conceptual connection between the items. To ensure that we had not missed a well-fitting model due to our decision to use parallel analysis to determine the number of factors to retain, we also tested a three-factor model based on the method of using Cattell's scree plot to determine the number of factors to retain. This model provided conceptual explanatory power but resulted in poor statistical model fit.

We evaluated the possibility that conducting separate analyses on the data from participants with high versus low levels of autistic traits would result in separate factor structures and better model fit for both subgroups. While we did find evidence that within our sample the factor structure was different between groups, rather than improving model fit, the fit indices indicated worse fit for both subgroups, and many items failed to load on to any factor.

The failure of this study and previous studies (Black, 2019; Olderbak et al., 2015) to identify a well-fitting factor structure for the RMET raises important questions about the reliability of this measure. Moreover, responses to particular RMET items in our study are of concern. As previously noted, the initial validity of the target mental states in the RMET was based on consensus. Baron-Cohen, Wheelwright, Hill et al. (2001) validated the individual RMET items in a combined sample of 225 participants consisting of Cambridge University students and members of the general public. The cutoff levels for consensus were "arbitrarily selected but with the aim of checking that a clear majority of the normal controls selected the target word and that this was selected at least twice as often as any foil" (Baron-Cohen, Wheelwright, Hill et. al., 2001, p. 244). In our study, eight items (22%) failed to pass one or both of the original criteria for retention in the test (i.e., less than 50% of participants selecting the target and more than 25% of participants selecting the same incorrect foil). While it is very rare for studies to provide a breakdown of target and foil response rates for individual RMET items, those studies that have done so regularly find that some RMET items do not meet these criteria (e.g., Eddy & Hansen, 2020; Olderbak et al., 2015; Prevost et al., 2014; Van Staden & Callaghan, 2021). We suggest that in future studies, researchers using the RMET should

report which items within their sample meet these criteria and provide raw data so that other researchers can explore further validity checks.

4.2. Do Levels of autistic traits or alexithymic traits correlate most strongly with RMET scores?

The third primary aim of this study was to assess the construct validity of the RMET by evaluating whether performance on the test is better explained by levels of autistic traits or alexithymic traits and whether performance correlates with another measure of ToM. Due to the limitations that we identified with the RMET in the factor analyses, we recommend caution drawing inferences from these analyses.

We predicted that both autistic traits as indexed by the AQ-28 and alexithymic traits as indexed by the TAS would negatively correlate with RMET scores, but that only the relationship between RMET scores and TAS scores would be statistically significant after controlling for TAS scores. This pattern of results would suggest that the poorer performance of individuals with high levels of autistic traits on the RMET more likely results from emotion recognition deficits associated with alexithymic traits than ToM deficits associated with autistic traits. We found that AQ-28 total scores did not correlate with RMET scores, even prior to controlling for TAS scores. This result is consistent with Oakley et al.'s (2016) finding that TAS scores are more predictive of RMET performance than AQ scores. However, looking at the subscales of the AQ-28 revealed a more complicated relationship. Two of the AQ-28 subscales, *imagination,* and *social skills*, did correlate with RMET scores, but in opposite directions. Unexpectedly, the *social skills* subscale positively correlated with RMET scores, indicating that RMET scores increased with an increase in autistic traits related to *social skills*. In contrast, the *imagination* subscale correlated negatively with RMET scores.

We also hypothesised that AQ-28 scores, but not TAS scores would correlate with the IMT ToM scores. However, we found the opposite result with TAS scores, but not AQ-28 scores, correlating with IMT ToM scores. One limitation of the IMT is that participants do not have access to the text when they are answering the questions. Thus, while the task provides both memory and ToM scores, both scores rely on memory ability, which may confound the results.

4.3. Does comfort viewing eyes impact RMET performance

We found that comfort viewing eye stimuli was positively correlated with RMET scores. That is, people who reported feeling more comfortable looking at the images of the eyes tended to score higher on the RMET. A detailed description of these analyses is provided in Part 2, section 2 of the supplementary materials. It is important to note that our measure of comfort viewing the eye stimuli was an ad-hoc, single item, self-report measure designed for this study. However, if researchers continue to use the RMET, our results indicate that further research should be conducted to determine the impact that comfort viewing eye stimuli has on RMET performance.

4.2 Conclusion

The RMET is a widely used measure of ToM ability in a variety of clinical and nonclinical populations. Despite being widely reported to be a well-validated tool, there is little empirical evidence to support this assertion, and converging evidence from the present study and other studies (Black, 2019; Olderbak et al., 2015; Kittel et al., 2021) raises considerable doubts about the reliability and validity of the RMET as a measure of ToM. Of particular concern, we failed to identify a well-fitting unidimensional or multidimensional factor model for the RMET suggesting that there are no discrete, consistent factors driving RMET response patterns.

Considering these issues, we suggest that the RMET may not be apt as a measure of ToM. The psychometric difficulties that come with the RMET indicate that past conclusions may need to be revised, and researchers should consider these issues before using the RMET or citing the conclusions of studies that use the RMET. The ongoing widespread use of the test despite evidence of its psychometric shortcomings is puzzling and may indicate insufficient attention being paid to measurement validity. Such “measurement schmeasurement” attitudes result in the use of measures without the provision of adequate evidence of their validity. This is a pervasive problem in psychology research, which threatens to undermine the validity of research conclusions (Flake & Fried 2020).

References

Bado, F. M. R., Rebustini, F., Jamieson, L., Cortellazzi, K. L., & Mialhe, F. L. (2018). Evaluation of the psychometric properties of the Brazilian version of the Oral Health Literacy Assessment in Spanish and development of a shortened form of the instrument. *PloS One, 13*(11), e0207989-e0207989. https://doi.org/10.1371/journal.pone.0207989

Bagby, R. M., Ayearst, L. E., Morariu, R. A., Watters, C., & Taylor, G. J. (2014). The internet administration version of the 20-Item Toronto Alexithymia Scale. *Psychological Assessment*, *26*(1), 16–22. https://doi.org/10.1037/a0034316

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23–32. https://doi.org/10.1016/0022-3999(94)90005-1

Benau, E. M., Wiatrowski, R., & Timko, C. A. (2020). Difficulties in emotion regulation, alexithymia, and social phobia are associated with disordered eating in male and female undergraduate athletes. *Frontiers in Psychology*, *11*, 1646–1646. https://doi.org/10.3389/fpsyg.2020.01646

Baron-Cohen, S. (1995). *Learning, development, and conceptual change. Mindblindness: An essay on autism and theory of mind.* MIT Press.

Baron-Cohen, S., Bowen, D. C., Holt, R. J., Allison, C., Auyeung, B., Lombardo, M. V., Smith, P., & 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), e0136521-e0136521. https://doi.org/10.1371/journal.pone.0136521

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, and Allied Disciplines*, *42*(2), 241–251. http://dx.doi.org/10.1111/1469-7610.00715

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5-17. https://doi.org/10.1023/A:1005653411471

Bergh, D. (2015). Sample size and chi-squared test of fit—A comparison between a random sample approach and a chi-square value adjustment method using Swedish adolescent data. In *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings* (pp. 197–211). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-47490-7_15

Bird, G., & Cook, R. (2013). Mixed emotions: the contribution of alexithymia to the emotional symptoms of autism. *Translational Psychiatry, 3*(7), e285 (2013). https://doi.org/10.1038/tp.2013.61

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., Vahip, S., Gonul, A. S., Akdeniz, F., Alkan, M., Ogut, M., & Eryavuz, A. (2005). Evidence for theory of mind deficits in euthymic patients with bipolar disorder. *Acta Psychiatrica Scandinavica*, *112*(2), 110–116. https://doi.org/10.1111/j.1600-0447.2005.00570.x

Botha, M., Hanlon, J., & Williams, G. L. (2021). Does language matter? Identity-first versus person first language use in autism research: A response to Vivanti. *Journal of Autism and Developmental Disorders*, 1–9. https://doi.org/10.1007/s10803-020-04858-w

Christmann, A., & Van Aelst, S. (2006). Robust estimation of Cronbach's alpha. *Journal of Multivariate Analysis, 97*(7), 1660-1674. https://doi.org/10.1016/j.jmva.2005.05.012

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319. https://doi.org/10.1037/1040-3590.7.3.309

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & politics, 6*(1). 205316801882217. https://doi.org/10.1177/2053168018822174

Deschrijver, E., & Palmer, C. (2020). Reframing social cognition: Relational versus representational mentalizing. *Psychological Bulletin, 146*(11), 941–969. https://doi.org/10.1037/bul0000302

Eddy, C. M., & Hansen, P. C. (2020). Predictors of performance on the Reading the Mind in the Eyes Test. *PloS One*, *15*(7), e0235529. https://doi.org/10.1371/journal.pone.0235529

Fazaeli, S. M., Amin Yazdi, S. A., Sharifi, S., Sobhani-Rad, D., & Ehsaei, M. R. (2018). Theory of mind in adults with traumatic brain injury. *Trauma Monthly*, *23*(4). https://doi.org/10.5812/traumamon.22022

Fernández-Abascal, E., Cabello, R., Fernández-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), 33–33. https://doi.org/10.1186/2040-2392-4-33

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

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*. https://doi.org/10.1177/2515245920951747

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), 102-118. https://doi.org/10.1037/arc0000067

Giordano, M., Licea-Haquet, G., Navarrete, E., Valles-Capetillo, E., Lizcano-Cortés, F., Carrillo-Peña, A., & Zamora-Ursulo, A. (2019). Comparison between the Short Story Task and the Reading the Mind in the Eyes Test for evaluating Theory of Mind: A replication report. *Cogent Psychology, 6*(1). https://doi.org/10.1080/23311908.2019.1634326

Gökçen, E., Frederickson, N., & Petrides, K. (2016). Theory of mind and executive control deficits in typically developing adults and adolescents with high levels of autism traits. *Journal of Autism and Developmental Disorders*, *46*(6), 2072–2087. https://doi.org/10.1007/s10803-016-2735-3

Golan, O., Sinai-Gavrilov, Y., & Baron-Cohen, S. (2015). The Cambridge Mindreading Face-Voice Battery for Children (CAM-C): complex emotion recognition in children with and without autism spectrum conditions. *Molecular Autism*, *6*(1), 22–22. https://doi.org/10.1186/s13229-015-0018-z

Goodboy, A. K., & Martin, M. M. (2020). Omega over alpha for reliability estimation of unidimensional communication measures. *Annals of the International Communication Association, 44*(4), 422-439. https://doi.org/10.1080/23808985.2020.1846135

Goretzko, D., Pham, T. T. H., & Bühner, M. (2019). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*. doi:10.1007/s12144-019-00300-2

Green, S. B., & Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: Coefficient alpha and omega coefficients. *Educational Measurement, Issues and Practice, 34*(4), 14-20. https://doi.org/10.1111/emip.12100

Hadjikhani, N., Åsberg Johnels, J., Zürcher, N. R., Lassalle, A., Guillon, Q., Hippolyte, L., Billstedt, E., Ward, N., Lemonnier, E., & Gillberg, C. (2017). Look me in the eyes: constraining gaze in the eye-region provokes abnormally high subcortical activation in autism. *Scientific Reports, 7*(1), 3163-3167. https://doi.org/10.1038/s41598-017-03378-5

Hair, J. F. Jr., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2014). *Multivariate data analysis* (7th ed.) Prentice-Hall.

Hallerbäck, M. U., Lugnegård, T., Hjärthag, 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., Stapleton, L.M., & Mueller, R.O. (Eds.). (2018). *The reviewer's guide to quantitative methods in the social sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9781315755649

Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154. https://doi.org/10.1007/bf02172093

Harkness, K., Sabbagh, M., Jacobson, J., Chowdrey, N., & Chen, T. (2005). Enhanced accuracy of mental state decoding in dysphoric college students. *Cognition and Emotion, 19*(7), 999 1025. https://doi.org/10.1080/02699930541000110

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hoekstra, R., Vinkhuyzen, A., Wheelwright, S., Bartels, M., Boomsma, D., Baron-Cohen, S., Posthuma, D., & Sluis, S. (2011). The construction and validation of an abridged version of the Autism-Spectrum Quotient (AQ-Short). *Journal of Autism and Developmental Disorders, 41*(5), 589-596. https://doi.org/10.1007/s10803-010-1073-0

Hu, L., & Bentler, P. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hurst, R., Mitchell, J, Kimbrel, N., Kwapil, T., & Nelson-Gray, R. (2007). Examination of the reliability and factor structure of the Autism Spectrum Quotient (AQ) in a non-clinical sample. *Personality and Individual Differences*, *43*(7), 1938–1949. https://doi.org/10.1016/j.paid.2007.06.012

Israelashvili, J., Sauter, D., & Fischer, A. (2020). Two facets of affective empathy: concern and

distress have opposite relationships to emotion recognition. *Cognition and Emotion*, *34*(6), 1112–1122. https://doi.org/10.1080/02699931.2020.1724893

Kallitsounaki, A., & Williams, D. (2020). Mentalising moderates the link between autism traits and current gender dysphoric features in primarily non-autistic, cisgender individuals. *Journal of Autism and Developmental Disorders, 50*(11), 4148–4157. https://doi.org/10.1007/s10803-020-04478-4

Khorashad, B., Baron-Cohen, S., Roshan, S., Kazemian, G., Khazai, M., Aghili, M., Talaei, L., & Afkhamizadeh, Z. (2015). The "Reading the Mind in the Eyes" test: Investigation of psychometric properties and test–retest reliability of the Persian version. *Journal of Autism and Developmental Disorders*, *45*(9), 2651–2666. https://doi.org/10.1007/s10803-015-2427-4

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology, 89*(2), 191-204. https://doi.org/10.1111/j.2044-8295.1998.tb02680.x

Kittel, A. F. D., Olderbak, S., & Wilhelm, O. (2021). 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 (Odessa, Fla.)*, 1073191121996469–1073191121996469. https://doi.org/10.1177/1073191121996469

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Koo, S. J., Kim, Y. J., Han, J. H., Seo, E., Park, H. Y., Bang, M., Park, J. Y., Lee, E., & An, S. K. (2021). "Reading the Mind in the Eyes Test": Translated and Korean versions. *Psychiatry Investigation*, *18*(4), 295–303. https://doi.org/10.30773/PI.2020.0289

Kuczynski, A. M., Kanter, J. W., & Robinaugh, D. J. (2020). Differential associations between interpersonal variables and quality-of-life in a sample of college students. *Quality of Life Research*, *29*(1), 127–139. https://doi.org/10.1007/s11136-019-02298-3

Launay, J., Pearce, E., Wlodarski, R., van Duijn, M., Carney, J., & Dunbar, R. I. M. (2015). Higher-order mentalising and executive functioning. *Personality and Individual Differences, 86*, 6-14. https://doi.org/10.1016/j.paid.2015.05.021

Livingston, L. A., Shah, P., Milner, V., & Happé, F. (2020). Quantifying compensatory strategies in adults with and without diagnosed autism. *Molecular Autism*, *11*(1), 15–15. https://doi.org/10.1186/s13229-019-0308-y

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320-341. https://doi.org/10.1207/s15328007sem1103_2

Meyer, J., & Shean, G. (2006). Social-cognitive functioning and schizotypal characteristics. *The Journal of Psychology, 140*(3), 199-207. https://doi.org/10.3200/JRLP.140.3.199-207

Muller, F., Simion, A., Reviriego, E., Galera, C., Mazaux, J.-M., Barat, M., & Joseph, P.-A. (2010). Exploring theory of mind after severe traumatic brain injury. *Cortex*, *46*(9), 1088–1099. https://doi.org/10.1016/j.cortex.2009.08.014

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion. *Journal of Abnormal Psychology*, *125*, 1–25. http://dx.doi.org/10.1037/abn0000182

Olderbak, S., Riggenmann, O., Wilhelm, O., & Doebler, P. (2021). Reliability generalization of tasks and recommendations for assessing the ability to perceive facial expressions of emotion. *Psychological Assessment*. https://doi.org/10.1037/pas0001030

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, 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*, 1–14. https://doi.org/10.3389/fpsyg.2015.01503

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks:

Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Ozturk, Y., Ozyurt, G., Turan, S., Mutlu, C., Tufan, A. E., & Akay, A. P. (2020). Association of theory of mind and empathy abilities in adolescents with social anxiety disorder. *Current Psychology (New Brunswick, N.J.)*. https://doi.org/10.1007/s12144-020-00707-2

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

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, (1978), 515–526. https://doi.org/10.1017/S0140525X00076512

Prevost, M., Carrier, M.-E., Chowne, G., Zelkowitz, 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

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior, 58*, 354–360. https://doi.org/10.1016/j.chb.2015.12.049

Revelle, W (2020). *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.0.7, https://CRAN.R-project.org/package=psych.

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

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism*, *6*(1), 2. https://doi.org/10.1186/2040-2392-6-2

Sherman, G. D., Lerner, J. S., Renshon, J., Ma-Kellams, C., & Joel, S. (2015). Perceiving others' feelings: The importance of personality and social structure. *Social Psychological & Personality Science*, *6*(5), 559–569. https://doi.org/10.1177/1948550614567358

Soker-Elimaliah, S., Jennings, C. A., Hashimi, M. M., Cassim, T. Z., Lehrfield, A., & Wagner, J. B. (2020). Autistic traits moderate relations between cardiac autonomic activity, interoceptive accuracy, and emotion processing in college students. *International Journal of Psychophysiology*, *155*, 118–126. https://doi.org/10.1016/j.ijpsycho.2020.04.005

Stiller, J., & Dunbar, R. I. M. (2007). Perspective-taking and memory capacity predict social network size. *Social networks, 29*(1), 93-104. https://doi.org/10.1016/j.socnet.2006.04.001

Stuart, N., Whitehouse, A., Palermo, R., Bothe, E., & Badcock, N. (2022). Eye gaze in autism spectrum disorder: A review of neural evidence for the eye avoidance hypothesis. *Journal of Autism and Developmental Disorders*. https://doi.org/10.1007/s10803-022-05443-z

Sulik, J., Ross, R. M., Balzan, R., & McKay, R. (2021). Delusion-like beliefs and data quality: Are classic cognitive biases artefacts of carelessness? https://doi.org/10.31234/osf.io/ntsve

Susana, L., Adoración, F., Ana, H., & Inés, T. (2017). The exploratory factor analysis of items: guided analysis based on empirical data and software. *Anales de Psicología, 33*(2), 417-432. https://doi.org/10.6018/analesps.33.2.270211

Trevisan, D. A., Roberts, N., Lin, C., & Birmingham, E. (2017). How do adults and teens with self-declared Autism Spectrum Disorder experience eye contact? A qualitative analysis of first-hand accounts. *PloS One, 12*(11), e0188446–e0188446. https://doi.org/10.1371/journal.pone.0188446

United States Census Bureau. (2019a). QuickFacts: United States. Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045219

United States Census Bureau. (2019b). Educational Attainment in the United States: 2019. Retrieved from https://www.census.gov/data/tables/2019/demo/educational-attainment/cps-detailed-tables.html

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*. https://doi.org/10.1007/s00426-021-01539-w

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

Wegner-Clemens, K., Rennig, J., & Beauchamp, M. S. (2020). A relationship between Autism-Spectrum Quotient and face viewing behavior in 98 participants. *PloS one, 15*(4), e0230866–e0230866. https://doi.org/10.1371/journal.pone.0230866

Zorowitz, S., Niv, Y., & Bennett, D. (2021). Inattentive responding can induce spurious associations between task behavior and symptom measures. https://doi.org/10.31234/osf.io/rynhk

Table 1

Factor Loadings for Three Factor EFA on the Full Sample

RMET item	Target	Factor 1 Thoughtful	Factor 2 Negative	Factor 3 Flirtatious
32	serious	0.514		
22	preoccupied	0.488		
24	pensive	0.486		
9	preoccupied	0.460		
17	doubtful	0.412		-0.408
15	contemplative	0.411		
16	thoughtful	0.396		
20	friendly	0.368		
27	cautious	0.343		
33	concerned	0.319		
28	interested	0.304		
4	insisting		0.490	
8	despondent		0.442	
26	hostile		0.440	
7	uneasy		0.411	
34	distrustful		0.407	
12	sceptical		0.361	
11	regretful		0.352	
35	nervous		0.310	
36	suspicious		0.306	
23	defiant		0.303	
30	flirtatious		0.498	0.550
21	fantasizing	0.327		0.462
3	desire			0.408
6	fantasizing			0.381
25	interested			0.351
1	playful			0.334

Note. Items were considered to load onto a factor if the factor loading was ≥ 0.3.

Table 2

Factor Loadings for Participants with Low AQ-28 Scores

RMET item	Target	Factor 1 Thoughtful	Factor 2 Negative	Factor 3 Flirtatious
16	thoughtful	0.603		
24	pensive	0.526		
29	reflective	0.414		
9	preoccupied	0.404		
22	preoccupied	0.401	0.343	
5	worried	0.358		
14	accusing	0.356		
28	interested	0.350		
13	anticipating	0.346		
34	distrustful	0.325		
12	sceptical	0.310		
4	insisting		0.579	
26	hostile		0.390	
36	suspicious		0.379	
8	despondent		0.363	
27	cautious		0.352	
11	regretful		0.343	
35	nervous		0.324	
3	desire			0.615
21	fantasizing			0.585
30	flirtatious	0.310		0.427
25	interested			0.417
31	confident			0.384

Note. Items were considered to load onto a factor if the factor loading was ≥ 0.3.

Table 3

Factor Loadings for Participants with High AQ-28 Scores

RMET item	Target	Factor 1	Factor 2 Negative	Factor 3 Thoughtful
30	flirtatious	0.807		
21	fantasizing	0.766	-0.405	
3	desire	0.509		
9	preoccupied	0.502		
32	serious	0.498		
25	interested	0.452		
13	anticipating	0.442		
26	hostile	0.441		
8	despondent	0.440		
29	reflective	0.437		
36	suspicious	0.428		
6	fantasizing	0.405		
1	playful	0.405		
31	confident	0.405		
16	thoughtful	0.400		
20	friendly	0.382		
15	contemplative	0.369		
18	decisive	0.367		
12	sceptical	0.362		
2	upset	0.358		
5	worried	0.329		
7	uneasy		0.430	0.329
34	distrustful	0.369	0.413	
4	insisting		0.326	
22	preoccupied			0.583
28	interested			0.528
24	pensive	0.357		0.374
17	doubtful			0.346

Note. Items were considered to load onto a factor if the factor loading was ≥ 0.3.

Table 4

SEM RMET Regression Results

Variables	*b*	SE(*b)*	β
Gender	0.690	0.290	0.064*
AQ-28 *imagination*	-0.159	0.042	-0.119***
AQ-28 *social skills*	0.110	0.036	0.106**
AQ-28 *switching*	0.121	0.075	0.051
AQ-28 *routine*	0.041	0.075	0.018
AQ-28 *numbers*	-0.026	0.046	-0.016
TAS *describe*	0.102	0.042	0.090*
TAS *identify*	-0.139	0.029	-0.170***
TAS *external*	-0.140	0.036	-0.117***
IMT memory	0.723	0.101	0.220***
IMT ToM	0.576	0.102	0.172***

*Note. * p < .05 ** p < .01 *** p < .001*