No evidence for the efficiency of the eye-tracking-based RMET version at detecting differences of mind reading abilities across psychological traits

Bertrand Beffara[1,2], Marina Veyrie[1,2], Laura Mauduit[1,2], Lara Bardi*[3], Irene Cristofori*[1,2]

[1]Institute of Cognitive Neuroscience Marc Jeannerod, CNRS / UMR 5229, 69500 Bron

[2]Université Claude Bernard, Lyon 1, 69100 Villeurbanne, France

[3]Gent University, Department of Psychology, Gent, Belgium

*Equal contribution

Corresponding author
Bertrand Beffara: bertrand.beffara@gmail.com

Abstract

The “Reading the Mind in the Eyes Test” (RMET) is one of the most used tests of theory of mind. Its principle is to match an emotion word to the corresponding face image. The performance at this test has been associated with multiple psychological variables including personality, loneliness and empathy. Recently, however, the validity of the RMET has been questioned. An alternative version of the test has been tested using eye-tracking (Russell et al., 2021) in addition to manual responses and was hypothesized to be more sensitive. Here, we put this hypothesis to the test by attempting to reproduce already-assessed correlational results between the performance at the classical RMET and the self-reported personality, loneliness and empathy, now using eye-gaze as an RMET performance index. Despite a marked eye-gaze bias towards the face image corresponding to the target word, the eye-gaze pattern correlated with none of the self-reported psychological variables. This result highlights the interest in using eye-tracking for theory of mind tests, while questioning the robustness of the association between psychological variables and RMET performance, and the validity of the RMET itself.

Keywords: Theory of Mind - Emotions – Personality – Loneliness – Empathy – RMET – Eye-tracking

Introduction

Social interactions can be defined as dynamic exchanges of information between individuals. They are key to psychological health in humans (Miller et al., 2009). To properly manage these interactions, humans use a set of cognitive abilities often referred to as “social cognition” (Frith, 2008; Seyfarth & Cheney, 2015). These abilities consist in correctly processing social signals and mainly comprise others’ recognition, others’ relationships understanding, social hierarchies understanding, others’ mental states, emotions, and goals understanding (i.e. theory of mind [ToM]) (Seyfarth & Cheney, 2015). Currently, various psychological tests/tasks are used to assess the participants’/patients’ social cognitive abilities (e.g. Bagby et al., 1994; Baron-Cohen et al., 1997; Baron-Cohen & Wheelwright, 2004; Doherty, 1997; Fett et al., 2011). One of the most influential test of the “theory of mind” component of social cognition is the Reading the Mind in the Eyes Test (RMET) (Baron-Cohen et al., 1997, 2001). This task is a test originally designed to evaluate the recognition of complex mental states expressed by human eyes. Typically, each trial consists in displaying a face picture accompanied by words describing different mental states. The participant’s goal is to select the word that best describes what the person in each picture is thinking or feeling. The test was showed to be able to detect subtle ToM deficits in high-function autistic individuals (Baron-Cohen et al., 1997, 2001).

Recently, the efficiency of the RMET at specifically measuring ToM has been challenged. Oakley et al. (2016) claimed that the poor performance of autism spectrum disorder (ASD) patients at the RMET could not be used as a validation of this test. Indeed, although ASD patients exhibit poor ToM abilities, ASD often co-occurs with alexithymia (difficulty in experiencing, identifying, and expressing emotions). On these grounds, the poor performance of ASD patients at the RMET could be explained by ToM impairments, emotion recognition difficulties, or both. Oakley et al. (2016) tested ASD patients and alexithymia-matched controls on the RMET. They found no significant performance difference between the two groups. However, when comparing RMET performance between alexithymic and non-alexithymic participants irrespective of ASD, they found the alexithymic participants exhibited poorer performance. They therefore suggested that the RMET was likely to test emotion recognition rather than ToM, for which the test has originally been designed (see also Kittel et al., 2022 for a meta-analysis). In addition, in an up-to-date systematic review gathering more than 1,400 research articles, Higgins et al. (2024) assessed the construct validity of the RMET and declared that this test’s validity was unsubstantiated because of no reliability in providing RMET’s validity evidence.

Russell et al. (2021) suggested that the response modality (i.e. selecting a word for the RMET) in emotion recognition tests could be a hurdle to their validity and proposed a computerized version of the RMET during which participants did not have to select a response-word by hand. Instead, on each trial they were presented with four face pictures accompanied by a single word describing an emotion (e.g. “happy”) while their spontaneous eye-gaze was recorded with an eye-tracking system. They observed that healthy participants looked at the face picture corresponding to the emotion-word reliably more often than at the three distractor face pictures. Patients with behavioural variant frontotemporal dementia – a disorder affecting social cognition skills – also exhibited this expected pattern of eye-gaze, but to a significantly lesser extent, suggesting that this version of the test was able to capture differences in emotion recognition between controls and the clinical group. Thanks to the analysis of a continuous measure, i.e. eye-tracking, the task could potentially be more sensitive to subtle differences in emotion recognition than the traditional version. In addition, neither experimental group in Russell et al. (2021) reached a ceiling (i.e. healthy participants' gaze was not 100% directed towards the target face picture) or floor (i.e. the gaze of patients with behavioural variant frontotemporal dementia was not randomly assigned to the face pictures) effect. Consequently, this modified version could also assess subtle differences in emotion recognition abilities. In the current study, we created an adapted version of the task designed by Russell and colleagues by adding a manual response, i.e., the recording of accuracy and reaction times. We reasoned that if this modified version of the RMET robustly assesses subtle differences in emotion recognition abilities, it should capture differences in psychological profiles in healthy participants that could not be systematically assessed using the classical RMET. To test this hypothesis, participants performed Russell et al.'s (2021) version of the RMET and correlated their task performance to psychological covariates that have already been associated with emotion recognition abilities, i.e. loneliness (e.g. Bosacki et al., 2020; Okruszek et al., 2021), personality (Richman & Unoka, 2015; see also Allen et al., 2017; Fertuck et al., 2009; Vonk et al., 2015) and empathy (e.g. Ibanez et al., 2013).

Methods

Participants

Forty-six participants were recruited for this study (18 men, mean age = 27 ± 8.1). They were all French speakers with normal or corrected vision and had a mean education time of 15.9 years (± 2.1). The study was approved by the Ethics committee of the University of Lyon, France (CER-UdL n° 2022-04-14-003).

Stimuli and procedure

The images stimuli were the images representing the eye region of faces expressing emotions selected for the original RMET (Baron-Cohen et al., 2001) and used also in Russell et al. (2021). Four pictures were presented on each trial following the same combination as in Russell et al. (2021). Each combination contained one target picture (“target”, matching the emotion expressed by the target word), one picture of the same valence as the target (“similar distractor”), and two pictures expressing completely different emotions (“different distractors”). On each trial, the four pictures were first presented, one in each quadrant of the screen, for 10s. In a second trial phase, the target word (i.e. the word allowing for the identification of the target picture) was centrally presented alone on the screen for 2s. During the last trial step, the central target word was presented along with the four pictures in each screen quadrant for 5s. Participants had to press one of four keys on a keyboard to select the image they thought best matched the target word. Response accuracy and time (RTs) were recorded (see **Figure 1**). There were one example and 20 testing trials. The trial order was pseudo-randomized. Before starting the experiment, all participants accessed a glossary defining all possible target words. During the entire task duration, continuous eye-tracking of both eyes was performed on I-motion using the Gazepoint GP3 eye-tracker. To minimize head movements, participants were leaned against a chin rest throughout the task. The display screen was 18″ and had a resolution of 1920 × 1080 pixels. The participant sat at 70cm from the screen. Before the task, a 9-point calibration was carried out on the Gazepoint software. If needed, a recalibration was performed until a good tracking accuracy was reached.

--------- Figure 1 almost here ----------

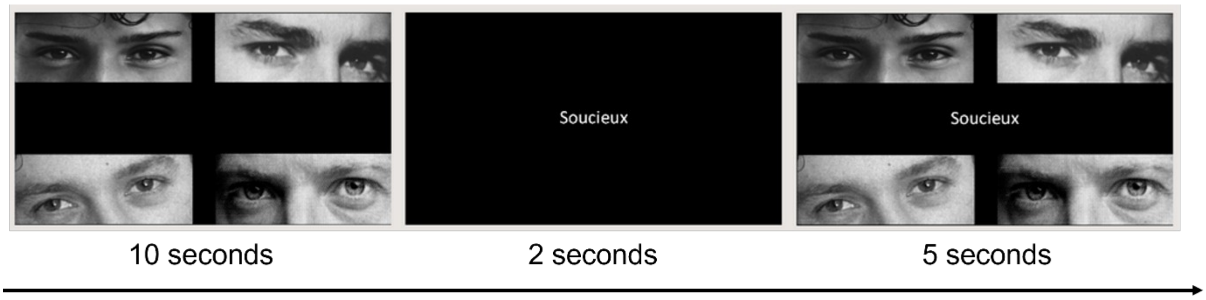

Figure 1: An example trial. The four pictures are first presented alone, one in each quadrant of the screen. In a second trial phase, the target word (i.e. the word allowing for the identification of the target picture) is centrally presented alone. During the last trial step, the central target word is presented along with the four pictures.

The main aim of the current study was to test to what extent the new version of the RMET proposed by Russell et al. (2021) is sensitive to differences in the psychological traits of healthy adult participants. To this end, after the main experiment, participants filled in questionnaires assessing their social skills. The first one was the interpersonal reactivity index (IRI; Davis, 1980; and see Gilet et al., 2013 for its validation in French; lower Cronbach's $\alpha > 0.69$), a multi-dimensional assessment of empathy which evaluates two cognitive components (fantasy and perspective-taking) and two affective components (empathic concern and personal distress). The second one was the Social and Emotional Loneliness Scale (SELSA; DiTommaso et al., 2004, 2007; Cronbach's $\alpha > 0.86$) measures three dimensions of loneliness: social loneliness, family loneliness, and romantic loneliness. The third one was the Revised NEO Personality Index (NEO PI-R; Costa Jr. & McCrae, 1997, Cronbach's $\alpha > 0.86$, see McCrae et al., 2011) which assesses five personality factors: Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C).

Statistical analyses

Areas of interests (AOI) for the analysis of fixation time were defined as each picture’s outline. Each participant’s dwell time within each AOI (i.e time spent looking at the 4 pictures) was measured, for each presented picture, during the first presentation of the pictures (10s) and during the second one (5s). In order to control for the presentation time, the percentage dwell time was computed as follows (see also Russell et al., 2021): *Dwell time (%) = (dwell time)/(presentation time) x 100*. The performance ("Dwell time change score") on each trial was measured as the difference between the percentage dwell time spent in the target AOI (target picture) before and after presentation of the "target emotion" word: *Dwell time change score = dwell time (%) post - dwell time (%) pre*. For each trial, the dwell time change scores for the distractor pictures (2 scores per trials) were averaged together. Therefore, this yielded a dwell time change score for each image type ("target", "similar distractor", "different distractor") for each participant. A "target focus score" was calculated by subtracting the dwell time change score corresponding to the different and similar distractor picture types to the target image type: *Target focus score = target dwell time change score – different distractor dwell time change score – similar distractor dwell time change score*.

The statistical analyses were performed on R Studio® using the *brms* package (Bürkner, 2017) suitable for Bayesian analyses. Flat priors were used for the parameter of interest (slope) and the gaussian family was used to fit the model. For all analyses, we reported the β parameter value of the model (slope) and the 95% credible interval.

The analyses of internal consistency for RT, accuracy and target focus score measures were performed using the *splithalf* R Studio® package (Parsons, 2021), which is particularly well-suited for RT and accuracy internal consistency analyses (Kahveci et al., 2024). Note that it may not be optimal for the estimation of the internal consistency of the target focus score measure, as it was not originally specifically designed for such use. However, to our knowledge this is currently the best-suited available approach (we used the “RT” option within the *splithalf* package, which corresponds to the closest available approximation of eye-gaze data characteristics). Briefly, the package relies on multiple (using random permutations) calculations of single internal consistency values using the Spearman-Brown index (see also Kahveci et al., 2024). For each permutation, data from each participant is split in two halves and the correlation between the two halves across participants is then used as the variable entered in the Spearman-Brown formula. The average internal consistency results from 6000 data permutations, therefore 6000 Spearman-Brown computations.

Processed data and analysis code are available on the Open Science Framework (https://osf.io/ze5v6/).

Results

Questionnaires internal consistency and descriptive statistics

The Spearman-Brown (S-B) index was used to estimate the internal reliability of all questionnaire measures. All subscales from the IRI, SELSA and NEO PI-R questionnaires were recalculated using the “split half” method. Accordingly, each subscale was computed twice: one using only the first half and one using only the second half of the items. For subscales comprising an uneven number of items, the first subscale computation was performed using one more item than the second computation. All subscales reached at least an internal consistency value of 0.62 (IRI: fantasy S-B index = 0.78, personal distress S-B index = 0.70, perspective taking S-B index = 0.84, empathic concern S-B index = 0.79; SELSA: social loneliness S-B index = 0.83, romantic loneliness S-B index = 0.92, family loneliness S-B index = 0.82; NEO PI-R: Neuroticism S-B index = 0.82, Extraversion S-B index = 0.76, Openness S-B index = 0.62, Agreeableness S-B index = 0.75, conscientiousness S-B index = 0.88).

	Mean (SD; Min - Max)
Loneliness scale	79.2 (15.3; 41 - 103)
Social loneliness	28.1 (5.4; 12 - 35)
Romantic loneliness	24.3 (9.1; 5 - 35)
Family loneliness	26.8 (6; 15 - 35)
Neo Pi-R scale	
Neuroticism	103.7 (21; 66 - 163)
Extraversion	106.9 (20.6; 53 - 150)
Openness	124.5 (21.8; 77 - 171)
Agreeableness	126.1 (20.5; 71 - 171)
Conscientiousness	116.2 (21; 71 - 156)
IRI scale	131.6 (22.2; 70 - 178)
Fantasy	34.1 (7.9; 16 - 49)
Personal distress	25.2 (8.2; 7 - 40)
Perspective taking	34.7 (8; 9 - 49)

Empathic concern	37.6 (7.06; 21 - 49)

Table1: Descriptive statistics of the questionnaire measures.

Internal consistency of implicit measures

Permutation-based split-half correlations (Kahveci et al., 2024) were used as internal consistency index for RT and accuracy variables. This yielded to a Spearman-Brown corrected reliability estimate of 0.78, 95% CI [0.68, 0.86] for RT, 0.26, 95% CI [-0.06, 0.53] for accuracy and 0.32, 95% CI [0.01, 0.58] for the target focus score (the latter result was taken from an analysis performed on 35 out of 46 participants due to missing data).

Dwell time change score analysis

Pairwise comparisons of dwell time change score across the three image types (see **Figure 2**) revealed that the mean dwell time change score for the target (mean = 8.65, standard deviation = 4.63) was higher than for both the similar images (mean = 0.47, standard deviation = 3.98) (Difference estimate = 8.19%; 95% CI = [6.53, 10.01]) and the distractor images (mean = -1.45, standard deviation = 3.24) (Difference estimate = 10.11%; 95% CI = [8.55, 11.70]). In addition, the mean dwell time change score for the similar images was higher than for the distractor images (β = 1.91; 95% CI = [0.50, 3.37]). All credibility intervals excluded zero, suggesting that the true difference estimate for each comparison is likely to be positive at the population level (see **Figure 2**).

--------- Figure 2 almost here ----------

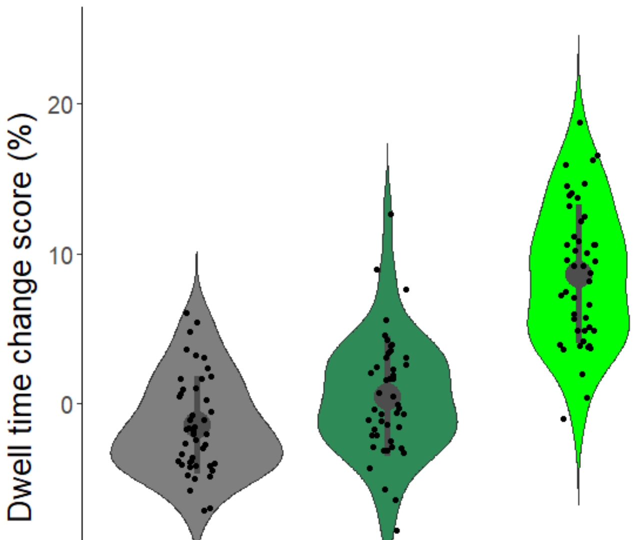

Figure 2: Violin plots of dwell time change scores across image types. Central indicators are means and ranges are standard deviations.

An analysis of the relationship between behavioural performance and pattern of eye-gaze revealed that the target focus score (i.e. to what extent the target image is visually explored relative to non-target images) positively correlates with the participants' accuracy (β = 0.38, 95% CI = [0.25, 0.51]). However, the corresponding analysis now regarding the relationship between RTs and target focus score did not reveal such correlation (β = 0.00, 95% CI = [-0.00, 0.00]). This suggests that a link exists between patterns of visually exploration and ability to perform the RMET, with no speed-accuracy trade-off. We further investigated this relationship in an exploratory analysis comparing target focus for correct vs. incorrect responses at the trial level. A mixed-model Bayesian analysis with participants entered as random effects allowing both varying intercepts and slopes revealed that the target focus was higher in trials for which participants gave a correct vs. incorrect response (β = 23.86, 95% CI = [-19.45, 28.21]) (see Figure 3). This further suggests that there exists a relationship between emotion recognition processes and visual perceptual/attentional processing of the target emotion.

--------- Figure 3 almost here ----------

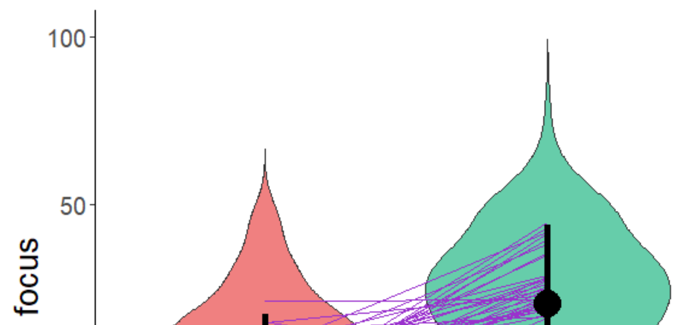

Figure 3: Target focus as a function of response correctness. Central indicators are means and ranges are standard deviations. Purple lines connect participant-specific means for correct vs. incorrect responses.

Associations between overall visual exploration patterns, behaviour, and psychological traits

Empathy, patterns of visual exploration and behavioural performance

Regarding eye-tracking data, the association between empathy and target dwell time change score is likely to be null or close to null (β = -0.04, 95% CI = [-0.10, 0.02]) and so is the relationship between empathy and target focus score (β = 0.05, 95% CI = [-0.05, 0.14]). Repeated across the IRI questionnaire subscales, correlation analyses with the target dwell time change score did not support the existence of non-null relationships (fantasy: β = -0.09, 95% CI = [-0.27, 0.08]; personal distress: β = -0.08, 95% CI = [-0.27, 0.11]; perspective taking: β = -0.04, 95% CI = [-0.21, 0.14]; empathic concern: β = -0.10, 95% CI = [-0.28, 0.10]). The same analyses performed with the target focus score showed no support for correlations between each IRI subscore a the target focus score (fantasy: β = 0.06, 95% CI = [-0.21, 0.34]; personal distress: β = 0.08, 95% CI = [-0.20, 0.34]; perspective taking: β = 0.16, 95% CI = [-0.13, 0.44]; empathic concern: β = 0.13, 95% CI = [-0.15, 0.42]).

Concerning the participants' behavioural responses, we found that the associations between the participants' response accuracy and empathy (β = 0.11, 95% CI = [-0.05, 0.28]) as well as between the participants' average RT and empathy (β = 1.13, 95% CI = [-14.77, 17.60]) are likely to be null or close to null. Analyses for each empathy questionnaire subscore showed that non-null correlations with behavioural performances are unprobable both for RT (fantasy: β = -11.04, 95% CI = [-57.94, 36.27]; personal distress: β = 1.35, 95% CI = [-44.18, 46.99]; perspective taking: β = -0.89, 95% CI = [-44.88, 41.31]; empathic concern: β = 20.71, 95% CI = [-28.98, 69.59]) and accuracy (fantasy: β = 0.24, 95% CI = [-0.22, 0.73]; personal distress: β = 0.11, 95% CI = [-0.40, 0.59]; perspective taking: β = 0.22, 95% CI = [-0.17, 0.71]; empathic concern: β = 0.40, 95% CI = [-0.13, 0.94]).

Overall, our results do not support an association between self-reported empathy and mind-reading abilities.

Loneliness, patterns of visual exploration and behavioural performance

At the level of eye-gaze patterns, we found that a non-null association both between loneliness and target dwell time change score (β = 0.02, 95% CI = [-0.08, 0.11]), and between loneliness and target focus score (β = 0.04, 95% CI = [-0.11, 0.18]) is unlikely. For completeness, we repeated these correlation analyses for the three subscales of the loneliness scale, i.e. social, romantic, and family loneliness. All non-null slopes regarding the association between the three subscales' scores and the target dwell time change score are unlikely (social loneliness: β = 0.09, 95% CI = [-0.16, 0.33]; romantic loneliness: β = -0.04, 95% CI = [-0.19, 0.10]; family loneliness: β = 0.15, 95% CI = [-0.08, 0.37]). The corresponding analyses, now targeting the association between the three subscales' scores and the target focus score did not support any positive/negative correlation (social loneliness: β = 0.36, 95% CI = [-0.01, 0.74]; romantic loneliness: β = -0.14, 95% CI = [-0.38, 0.10]; family loneliness: β = 0.33, 95% CI = [-0.02, 0.67]).

At the level of the behavioural responses, the participants' accuracy and any dimension of loneliness do not correlate (social loneliness: β = 0.24, 95% CI = [-0.43, 0.90]; romantic loneliness: β = -0.10, 95% CI = [-0.50, 0.30]; family loneliness: β = 0.02, 95% CI = [-0.62, 0.64]). The corresponding RTs analysis did not support the existence of a relation between any dimension of loneliness and RTs (social loneliness: β = -10.78, 95% CI = [-74.07, 50.96]; romantic loneliness: β = -17.78, 95% CI = [-54.96, 20.61]; family loneliness: β = 38.24, 95% CI = [-15.54, 94.89]).

Together, these results do not support an association between loneliness, patterns of visual activity during the task, and mind-reading abilities.

Personality, patterns of visual exploration and behavioural performance

We next performed correlation analyses between each personality dimension and the two eye-tracking indexes (target dwell time change score and target focus score), and between each personality dimension and the two responses metrics (accuracy and RTs).

The associations between all personality dimensions and target dwell time change score are likely to be null or close to null (Neuroticism: β = 0.00, 95% CI = [-0.06, 0.07]; Extraversion: β = 0.01, 95% CI = [-0.06, 0.08]; Openness: β = 0.01, 95% CI = [-0.05, 0.08]; Agreeableness: β = -0.04, 95% CI = [-0.11, 0.02]; Conscientiousness: β = -0.02, 95% CI = [-0.09, 0.05]). The analyses revealed the same pattern of results, now when correlating target focus scores with personality scores (Neuroticism: β = 0.04, 95% CI = [-0.06, 0.14]; Extraversion: β = -0.03, 95% CI = [-0.13, 0.08]; Openness: β = 0.04, 95% CI = [-0.05, 0.14];

Agreeableness: β = 0.08, 95% CI = [-0.02, 0.18]; Conscientiousness: β = 0.06, 95% CI = [-0.04, 0.16]).

Regarding the link between behavioural responses and personality components, all correlation slopes are unlikely to differ from zero let it be when using response accuracy (Neuroticism: β = 0.06, 95% CI = [-0.13, 0.23]; Extraversion: β = -0.11, 95% CI = [-0.30, 0.08]; Openness: β = 0.11, 95% CI = [-0.06, 0.29]; Agreeableness: β = 0.10, 95% CI = [-0.09, 0.30]; Conscientiousness: β = 0.06, 95% CI = [-0.11, 0.24]) or RTs (Neuroticism: β = 0.33, 95% CI = [-17.40, 17.27]; Extraversion: β = -7.22, 95% CI = [-23.40, 10.51]; Openness: β = -5.80, 95% CI = [-21.71, 10.99]; Agreeableness: β = 8.89, 95% CI = [-8.13, 26.03]; Conscientiousness: β = 0.06, 95% CI = [-0.11, 0.24]) as the dependent variable.

Together, these results do not support any association between personality traits and mind-reading abilities.

Discussion

In this study, we aimed to evaluate the efficiency of the eye-tracking-based version of the RMET proposed by Russell et al. (2021) to detect changes in mind-reading abilities across different psychological features that have been associated with emotion recognition (Bosacki et al., 2020; Okruszek et al., 2021 for loneliness-emotion recognition associations; Allen et al., 2017 for personality-emotion recognition associations; and Ibanez et al., 2013 for empathy-emotion recognition associations). The RMET recently underwent intense criticism (see Higgins et al., 2024 for a recent review) and Russell et al.'s eyetracking-based version of this test (2021) came as a possible methodological solution as it was hypothesized to be more sensitive and specific to differences in mind-reading abilities as an implicit quantitative index.

While this hypothesis was tested on healthy participants vs. frontotemporal dementia patients in Russell et al. (2021), here our main result is that this eye-tracking-based version of the RMET does not appear as a clearly efficient methodology at detecting differences in emotion recognition/mind-reading abilities across more subtle differences in psychological traits. Indeed, no likely correlations were found between the multiple psychological traits tested here and mind-reading/emotion recognition abilities as assessed by gaze patterns.

First, we deem it unlikely that our results did not reveal likely positive/negative correlations because of a lack of power of the eye-tracking measure *per see* to detect differences in the domain of social cognition because this type of measure is extensively used in studies investigating patterns of visual activity in the presence of social visual stimuli (see e.g. Birmingham et al., 2008, 2009; Böckler et al., 2014; Flechsenhar & Gamer, 2017; Martinez-Cedillo & Foulsham, 2024). Moreover, our analyses showed that 1) the mean dwell time change is likely to be higher for the target image than for the non-target images, suggesting that the eye-tracking measure correctly detects changes in mind-reading/emotion recognition dynamics (i.e. the visual exploration of the target image shows the highest increase after vs. before the target word presentation, compared to distractor and non-target images) and 2) the target focus score likely positively correlates with response accuracy, suggesting that eye-gaze patterns and task performance are related.

Secondly, this latter argument also supports the fact that the current version of the RMET actually measures mind-reading/emotion recognition: the target dwell time raises specifically when a matching between a word describing an emotion and the corresponding target image is possible (i.e. last trial phase, see Figure 1). This is an advantage of the current task version because the classical versions of the RMET (Baron-Cohen et al., 2001) did not comprise an implicit behavioural measure of the visual exploration of/attention to each image. This is an important consideration, because Higgins et al.'s review (2024) points out possible confounds in initial works that compared RMET performance in autistic vs. non-autistic participants (Baron-Cohen et al., 1997, 2001) as a proof-of-concept: the difference between autistic vs. non-autistic participants could not only rely on ToM abilities, but also e.g. on visual discomfort. Here our control analyses suggest, on the contrary, that the task accurately generates cognitive matching in an emotion recognition context. From an attentional point of view, data supporting a spatial bias towards a specific region of interest often reflect a matching between task instructions (see e.g. the notion of "attentional template" in Chelazzi et al., 1998) and the current visual stimuli. In visual scenes involving social stimuli, it has been shown in recent studies that task instructions modify overt attention towards these social stimuli (Flechsenhar & Gamer, 2017; Martinez-Cedillo & Foulsham, 2024). In the context of the current study, we interpret the mean dwell time change results similarly: the evolution of the content of the “attentional template” over time (before vs. after the target word presentation) drives the boost of spatial overt attention towards the target image. In addition, the target focus score is likely higher when the correct (vs. incorrect) response is selected by the participants. This probably translates that implicit perceptual/attentional cognitive processes are at the play in this version of the RMET. These results provide evidence for an efficient measure of the cognitive matching of emotional information. On these grounds, the current eye-tracking-based version of the RMET (see also Russell et al., 2021) is a promising tool for the investigation of emotion recognition abilities that, to our knowledge, had never been tested on non-elderly participants (cf. the study limitation discussed in Russell et al., 2021).

We, however, question the ability of this current RMET version to specifically and accurately detect subtle differences in mind-reading abilities across psychological profiles. First, it is unclear if the stimuli used for the RMET allow to capture ToM or emotion recognition abilities. Indeed, Oakley et al. (2016) had autistic-patients vs. controls performing the RMET and did not find a significant performance difference. However, now comparing alexithymic (i.e. poor abilities to recognize one's own emotions) vs. non-alexithymic participants, they found that alexithymic participants performed worse than non-alexithymic participants. They concluded that the RMET measures emotion recognition rather than ToM, hypothetically because of the emotional nature of the visual stimuli. Secondly, whatever the RMET measures, in the current study we highlighted that the participants' performance at this test is not likely to be associated with differences in the psychological traits we measured (see also Hendel & Brysbaert, 2024, who found that objective emotion recognition test primarily reflect intelligence instead of social-emotional abilities).

Overall, the correlational analyses may appear as a "null result" from the current study, which comes in contradiction with multiple larger sample size studies reporting significant correlations between ToM and loneliness (Bosacki et al., 2020; Okruszek et al., 2021), personality (Allen et al., 2017; Richman & Unoka, 2015) or empathy (Ibanez et al., 2013). We acknowledge that the current study's sample size could not allow us to detect small effects, and therefore only questions relationships of medium to large sizes between psychological traits and RMET measures. However, our primary objective was to test eye-tracking as a mind-reading assessment method more sensitive to differences in psychological traits than the traditional approach (used in studies similar to the current one, see Bosacki et al., 2020; Ibanez et al., 2013; Okruszek et al., 2021). Accordingly, we would have expected inflated effect sizes compared to the ones previously reported, and thus higher probability to detect them with the current sample size. On the contrary, none of the multiple correlation analyses performed here revealed probable correlations between eye-tracking-based RMET performance and psychological traits. Together with the current questioning of the RMET validity (see Higgins et al., 2024) and given that the field of psychology suffers from a bias towards publishing “positive” (vs. “null”) results (Nosek et al., 2022), our results come as further evidence 1) putting a note of caution regarding the RMET efficiency at detecting ToM/emotion recognition abilities and/or 2) questioning the sensitivity of the RMET to differences in psychological traits. This should also push into considering the creation of new validated ToM/emotion recognition tests. For example, Franca et al. (2023) identified an RMET weakness in that – if valid – it tests the recognition of *complex* emotions, which is subject to bias such as the participants’ verbal proficiency. Accordingly, in their study, and based on the RMET features, they designed a new test of recognition of *basic* emotions that is hypothesized to be less sensitive to unexpected/unbalanced participants’ psychological skills.

As a limitation, one possible explanation of the null results observed in the current study could regard the score ranges obtained for each questionnaire measure. Indeed, a limited score range could prevent from observing a non-null correlation even if it would appear as such given a wider distribution of the questionnaire scores. To estimate this possible bias, we numerically compared the questionnaire scores obtained in the current study (cf. Table 1) vs. in the existing literature. The current NEO PI-R and IRI scores are consistent with existing data (see Costa Jr. & McCrae, 1997; Gilet et al., 2013, respectively). However, the participants in the current study reported less mean loneliness (i.e. higher SELSA scores), and the variability was reduced, compared to what was expected from the literature (DiTommaso et al., 2004). This means that a ceiling effect possibly caused the observation of null correlations between loneliness and patterns of eye-gaze performed the RMET. We, however, deem it unlikely since all other correlations (cf. relationships between eye-gaze patterns and IRI/NEO PI-R) were found to be likely to be null.

In conclusion, we found no evidence that the eye-tracking-based version of the RMET initially proposed by Russell et al. (2021) can come as a solution to support the universal RMET validity as a psychological test of mind-reading and/or emotion recognition. However, this version allows for robust controls of the implicit cognitive states of the participants and could be used as a reference method for the setting of a more valid measure of mind-reading abilities following Higgins et al.'s (2024) main recommendations, including a clear assessment of which specific aspects of social cognition are measured by the RMET.

Funding

This work was supported by the Agence Nationale de la Recherche (Grant number ANR-20-CE28-0006 to Lara Bardi, acronym: IMPTOM), and the Claude Bernard University of Lyon 1 (SENS funding; R05SENS_SENS23CRI), the associations Liv & Lumière, LEEM (Les entreprises du médicament), Neurodis to Irene Cristofori.

References

Allen, T. A., Rueter, A. R., Abram, S. V., Brown, J. S., & Deyoung, C. G. (2017). Personality and Neural Correlates of Mentalizing Ability. *European Journal of Personality*, *31*(6), 599–613. https://doi.org/10.1002/per.2133

Bagby, R. M., Parker, J. D. A., & 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

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. (2004). The Empathy Quotient: An Investigation of Adults with Asperger Syndrome or High Functioning Autism, and Normal Sex Differences. *Journal of Autism and Developmental Disorders*, *34*(2), 163–175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

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/10.1111/1469-7610.00715

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Social Attention and Real-World Scenes: The Roles of Action, Competition and Social Content. *Quarterly Journal of Experimental Psychology*, *61*(7), 986–998. https://doi.org/10.1080/17470210701410375

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision Research*, *49*(24), 2992–3000. https://doi.org/10.1016/j.visres.2009.09.014

Böckler, A., Hömke, P., & Sebanz, N. (2014). Invisible Man: Exclusion From Shared Attention Affects Gaze Behavior and Self-Reports. *Social Psychological and Personality Science*, *5*(2), 140–148. https://doi.org/10.1177/1948550613488951

Bosacki, S., Moreira, F. P., Sitnik, V., Andrews, K., & Talwar, V. (2020). Theory of Mind, Self-Knowledge, and Perceptions of Loneliness in Emerging Adolescents. *The Journal of Genetic Psychology*, *181*(1), 14–31. https://doi.org/10.1080/00221325.2019.1687418

Bürkner, P.-C. (2017). **brms**: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Chelazzi, L., Duncan, J., Miller, E. K., & Desimone, R. (1998). Responses of Neurons in Inferior Temporal Cortex During Memory-Guided Visual Search. *Journal of Neurophysiology*, *80*(6), 2918–2940. https://doi.org/10.1152/jn.1998.80.6.2918

Costa Jr., P. T., & McCrae, R. R. (1997). Stability and Change in Personality Assessment: The Revised NEO Personality Inventory in the Year 2000. *Journal of Personality Assessment*, *68*(1), 86–94. https://doi.org/10.1207/s15327752jpa6801_7

Davis, M. H. (1980). *Interpersonal Reactivity Index* [Dataset]. https://doi.org/10.1037/t01093-000

DiTommaso, E., Brannen, C., & Best, L. A. (2004). Measurement and Validity Characteristics of the Short Version of the Social and Emotional Loneliness Scale for Adults. *Educational and Psychological Measurement*, *64*(1), 99–119. https://doi.org/10.1177/0013164403258450

DiTommaso, E., Turbide, J., Poulin, C., & Robinson, B. (2007). L’ÉCHELLE DE SOLITUDE SOCIALE ET ÉMOTIONNELLE (ÉSSÉ): A FRENCH-CANADIAN ADAPTATION OF THE SOCIAL AND EMOTIONAL LONELINESS SCALE FOR

ADULTS. *Social Behavior and Personality: An International Journal*, *35*(3), 339–350. https://doi.org/10.2224/sbp.2007.35.3.339

Doherty, R. W. (1997). The emotional contagion scale: A measure of individual differences. *Journal of Nonverbal Behavior*, *21*(2), 131–154. https://doi.org/10.1023/A:1024956003661

Fertuck, E. A., Jekal, A., Song, I., Wyman, B., Morris, M. C., Wilson, S. T., Brodsky, B. S., & Stanley, B. (2009). Enhanced 'Reading the Mind in the Eyes' in borderline personality disorder compared to healthy controls. *Psychological Medicine*, *39*(12), 1979–1988. https://doi.org/10.1017/S003329170900600X

Fett, A.-K. J., Viechtbauer, W., Dominguez, M.-G., Penn, D. L., Van Os, J., & Krabbendam, L. (2011). The relationship between neurocognition and social cognition with functional outcomes in schizophrenia: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*(3), 573–588. https://doi.org/10.1016/j.neubiorev.2010.07.001

Flechsenhar, A. F., & Gamer, M. (2017). Top-down influence on gaze patterns in the presence of social features. *PLOS ONE*, *12*(8), e0183799. https://doi.org/10.1371/journal.pone.0183799

Franca, M., Bolognini, N., & Brysbaert, M. (2023). Seeing emotions in the eyes: A validated test to study individual differences in the perception of basic emotions. *Cognitive Research: Principles and Implications*, *8*(1), 67. https://doi.org/10.1186/s41235-023-00521-x

Frith, C. D. (2008). Social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1499), 2033–2039. https://doi.org/10.1098/rstb.2008.0005

Gilet, A.-L., Mella, N., Studer, J., Grühn, D., & Labouvie-Vief, G. (2013). Assessing dispositional empathy in adults: A French validation of the Interpersonal Reactivity Index (IRI). *Canadian Journal of Behavioural Science / Revue Canadienne Des Sciences Du Comportement*, *45*(1), 42–48. https://doi.org/10.1037/a0030425

Hendel, E., & Brysbaert, M. (2024). *Towards understanding the low correlation between subjective and performance-based measures of emotion perception: Is one measure better than the other?* https://doi.org/10.31219/osf.io/mcfzt

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

Ibanez, A., Huepe, D., Gempp, R., Gutiérrez, V., Rivera-Rei, A., & Toledo, M. I. (2013). Empathy, sex and fluid intelligence as predictors of theory of mind. *Personality and Individual Differences*, *54*(5), 616–621. https://doi.org/10.1016/j.paid.2012.11.022

Kahveci, S., Bathke, A. C., & Blechert, J. (2024). Reaction-time task reliability is more accurately computed with permutation-based split-half correlations than with Cronbach's alpha. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02597-y

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

Martinez-Cedillo, A. P., & Foulsham, T. (2024). Don’t look now! Social elements are harder to avoid during scene viewing. *Vision Research*, *216*, 108356. https://doi.org/10.1016/j.visres.2023.108356

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal Consistency, Retest Reliability, and Their Implications for Personality Scale Validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

Miller, G., Chen, E., & Cole, S. W. (2009). Health Psychology: Developing Biologically Plausible Models Linking the Social World and Physical Health. *Annual Review of Psychology*, *60*(1), 501–524. https://doi.org/10.1146/annurev.psych.60.110707.163551

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*, 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

Okruszek, Ł., Piejka, A., Krawczyk, M., Schudy, A., Wiśniewska, M., Żurek, K., & Pinkham, A. (2021). Owner of a lonely mind? Social cognitive capacity is associated with objective, but not perceived social isolation in healthy individuals. *Journal of Research in Personality*, *93*, 104103. https://doi.org/10.1016/j.jrp.2021.104103

Parsons, S. (2021). splithalf: Robust estimates of split half reliability. *Journal of Open Source Software*, *6*(60), 3041. https://doi.org/10.21105/joss.03041

Richman, M. J., & Unoka, Z. (2015). Mental state decoding impairment in major depression and borderline personality disorder: Meta-analysis. *British Journal of Psychiatry*, *207*(6), 483–489. https://doi.org/10.1192/bjp.bp.114.152108

Russell, L. L., Greaves, C. V., Convery, R. S., Nicholas, J., Warren, J. D., Kaski, D., & Rohrer, J. D. (2021). Novel instructionless eye tracking tasks identify emotion recognition deficits in frontotemporal dementia. *Alzheimer’s Research & Therapy*, *13*(1), 39. https://doi.org/10.1186/s13195-021-00775-x

Seyfarth, R. M., & Cheney, D. L. (2015). Social cognition. *Animal Behaviour*, *103*, 191–202. https://doi.org/10.1016/j.anbehav.2015.01.030

Vonk, J., Zeigler-Hill, V., Ewing, D., Mercer, S., & Noser, A. E. (2015). Mindreading in the dark: Dark personality features and theory of mind.

Personality and Individual Differences, *87*, 50–54. https://doi.org/10.1016/j.paid.2015.07.025