OPEN MIND

Discoveries in Cognitive Science

an open access journal

Check for updates

Citation: González, R., López, J. I., & Abal, F. J. P. (2025). An Item Response Theory Analysis of the "Reading the Mind in the Eyes Test" In An Argentine Sample. *Open Mind: Discoveries in Cognitive Science*, *9*, 1121–1137. https://doi.org/10.1162/opmi.a.11

DOI: https://doi.org/10.1162/opmi.a.11

Received: 26 December 2024
Accepted: 1 June 2025

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Rocio González
mgrociogonzalez@gmail.com

Copyright: © 2025 Rocio González, Juan Ignacio López, and Facundo Juan Pablo Abal. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

REPORT

An Item Response Theory Analysis of the "Reading the Mind in the Eyes Test" In An Argentine Sample

Rocio González[1], Juan Ignacio López[2], and Facundo Juan Pablo Abal[3]

[1]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad Siglo 21, Instituto de Ciencias para la Familia, Universidad Austral. Universidad Argentina de la Empresa (UADE)
[2]Universidad Argentina de la Empresa (UADE)
[3]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad de Buenos Aires

Keywords: reading the mind in the eyes test, theory of mind, item response theory, argentine adults

ABSTRACT

In the field of affective neuroscience, the "Reading the Mind in the Eyes Test" (RMET) is regarded as the gold standard for assessing Theory of Mind (ToM) in general and clinical populations. Despite its widespread use and acceptance, there has been no comprehensive evaluation of the psychometric properties of the Spanish version of the test, particularly concerning its internal structure and dimensionality, especially with regard to the application of Item Response Theory (IRT). Therefore, the aim of this study is to examine the items of the Argentine version of the RMET through an IRT approach to explore its effectiveness in measuring ToM in neurotypical adult populations within Argentina. A sample of 899 adults from Buenos Aires, Argentina (34.3% men, 65.7% women) aged between 18 and 82 (M = 32.95, SD = 13.35) completed a sociodemographic questionnaire and the RMET. Based on a preliminary assessment of the data's suitability for factor analysis, including inspection of item-level sampling adequacy, we analysed a reduced 13-item version of the test. While these items were consistent with a unidimensional structure, the main factor explained only 21.2% of the variance, and the factor loadings were relatively low. The analysis revealed weak and questionable evidence for the dimensionality of the construct, with retained items showing medium-low difficulty and moderate discrimination. Additionally, the overall reliability estimates were insufficient. Therefore, the results of this study do not support the psychometric properties of the argentine version of RMET and contribute to increasing awareness of the critical need to thoroughly evaluate its validity and reliability.

INTRODUCTION

Social cognition, defined as a set of neurocognitive processes—both conscious and unconscious—that underlie social interactions (Atenas et al., 2019), is an important variable in the field of affective neurosciences. This is considered a multi-level or multi-layered phenomenon that includes Theory of Mind (ToM), or mentalization, which refers to cognitive representations of the mental states of others (Kliemann et al., 2021; Schurz et al., 2021). In essence, ToM refers to the ability of individuals to recognize that others have a mind, whose content is different from their own, which allows them to explain and predict the intentions, desires, and behaviors of others more accurately (López et al., 2014). Furthermore, the more

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

developed this skill is, the easier it will be for them to adapt to social groups; while its deficiency can be highly disabling (Atenas et al., 2019; Roman et al., 2012).

Social cognition has gained considerable attention in recent years, and it is included as one of the six core neurocognitive domains in the latest edition of the Diagnostic and Statistical Manual (DSM-5; APA, 2013). Furthermore, it plays a significant role in conditions such as Autism Spectrum Disorder, Schizophrenia, and Attention Deficit Disorder, where impairments in attributing intentionality to others are commonly found (Alvarez et al., 2023; Deng et al., 2024; Johnson et al., 2022; López & Richaud de Minzi, 2020; Peñuelas-Calvo et al., 2019; Stafford et al., 2023).

"The Reading the Mind in the Eyes Test" (RMET), developed by Baron-Cohen et al. (1997), is the gold standard task used to assess ToM in both general and clinical populations. The test comprises a series of facial images—focused on the eye region—in which participants must identify the portrayed mental state. In recent years, the RMET has been widely adopted as a key measure of ToM (Kittel et al., 2022; Stafford et al., 2023), specifically to measure the capacity to infer mental states (i.e., what another person is thinking or feeling) based on immediately observable cues (such as facial expressions, specifically the eyes) (Black, 2018; Roman et al., 2012).

The RMET has been translated into more than 30 languages and has undergone multiple validation studies worldwide (López & Richaud de Minzi, 2020). Furthermore, the test not only differentiates non-clinical from clinical populations, but has also been used in high-risk clinical contexts, such as awake brain surgery, to preserve areas involved in social cognition (Chung & Yoon, 2019; Higgins, Kaplan, et al., 2024). Moreover, the test is formally included in the Research Domain Criteria (RDoC) framework of the U.S. National Institute of Mental Health (NIMH), where it is recognized as the most robust tool for assessing the ability to infer mental states (NIMH, 2016).

However, despite its widespread popularity and use, a large-scale evaluation of the test's psychometric properties has yet to be conducted (Kittel et al., 2022). To date, studies have carried out the necessary translations into each language and assessed the score distributions and internal consistency of the instrument using Cronbach's alpha or test-retest reliability. Internal consistency values reported are often between poor and deficient (roughly within a range of .30 to .70), whereas the test-retest reliability of the measure is typically deemed acceptable (Higgins, Kaplan, et al., 2024; Olderbak et al., 2015). However, few studies have examined the internal structure of the test, and to date, this remains unclear (Higgins, Kaplan, et al., 2024). Additionally, there are a limited number of studies that have applied Item Response Theory (IRT): To our knowledge: Black, 2018; Cohen et al., 2019; Miguel et al., 2017; Preti et al., 2017).

Normative data for the test in Argentina have been presented for neurotypical adults from Buenos Aires City and Greater Buenos Aires (Roman et al., 2012). Furthermore, a recently adapted 18-item shortened version of the test for the Argentine population has been introduced, although it reports medium-low internal consistency with a Cronbach's alpha coefficient of .64 (López & Richaud de Minzi, 2020). Although this study (like most studies using the RMET; see Deng et al., 2024; Johnson et al., 2022; Peñuelas-Calvo et al., 2019) found statistically significant differences in test performance between the neurotypical and clinical samples (like autism spectrum disorders or schizophrenia), no ROC curve analyses have been conducted to evaluate the test's sensitivity and diagnostic accuracy. Moreover, neither local studies nor most international studies have examined the psychometric properties related to the test's internal structure and dimensionality.

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

Therefore, the aim of this study is to examine the items of the full Argentine version of the RMET using IRT to explore how the RMET might measure the latent construct of Theory of Mind in the neurotypical adult local population. In this way, the study seeks to address issues related to the quality of the RMET by evaluating its validity and reliability through advanced psychometric methods.

METHOD

Study Design & Participants

The present study adopts a psychometric instrumental design, according to the classification criteria proposed by Ato et al. (2013).

The sample included 899 adults aged between 18 and 82 years (M = 32.95; *SD* = 13.35) from Buenos Aires (Argentina), selected through non-probability purposive sampling, of whom 65.7% were women and 34.3% were men. Furthermore, 27.5% were married, 65.4% were single, 6.5% were divorced, and 0.7% were widowed, while most participants (70.6%) did not have children. As for educational background, 37.0% held a completed university degree, 62.2% had completed secondary education, and 0.8% had only completed primary education. Regarding self-perceived socioeconomic status, 4.2% considered themselves low, 26.0% as lower-middle, 56.0% as middle, 12.6% as upper-middle, and 1.2% as high.

Materials

Sociodemographic Questionnaire. A self-administered questionnaire was developed specifically to collect sociodemographic information from the participants, such as sex, age, marital status, highest level of education achieved, and self-perceived socioeconomic status.

The Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001), as adapted for the Argentine population (Roman et al., 2012). The RMET consists of 36 black-and-white photographs depicting only the eye region of faces, representing specific mental states. Each image is standardized to a size of 15 cm × 6 cm, displaying a rectangular area edited such that the eye region is visible from just above the eyebrow down to midway along the bridge of the nose. Additionally, for each photograph, four adjectives or terms are provided to describe the mental state of the individual in the image (one target and three distractors), and the participant is asked to choose the word that most accurately reflects the feelings or mental state of the subject in the picture. Additionally, the test is considered a measure of maximum performance, where each item scored as '1' for a correct response and '0' for an incorrect one, with a total score range from 0 to 36. The Spanish translation documents are freely accessible on the Autism Research Centre website: https://www.autismresearchcentre.com/research/tests/.

Procedure

All participants were contacted by email, informed about the purpose of the study, and invited to participate. Additionally, their participation was entirely voluntary, with no feedback or compensation provided, assuring data confidentiality and the right to withdraw from the study at any time. All these considerations were part of the informed consent form, which participants digitally signed. The test was completed by participants in a self-administered format using the Google Forms platform. First, the participant reads the instructions, followed by the practice item. Then, each photograph is presented (one at a time) with the four response options. As the participant chooses the answer they consider correct, they advance to the next photograph until they complete the test. The order of the RMET items was not randomized

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

across participants. The estimated time it takes to complete the entire test is approximately 15 minutes, although administration was self-managed. During the development of the study, the ethical principles of research with human beings were respected, ensuring the necessary conditions to protect the anonymity and confidentiality of the data. Likewise, the research and the protocol were approved by the Bioethics Committee of the National University of Mar del Plata, registered in the Provincial Registry of Research Ethics Committees, dependent on the Central Research Ethics Committee (Ministry of Health of the Province of Buenos Aires, Argentina).

Data Analysis

Responses from the participants were dichotomously coded based on the scoring key. An exploratory factor analysis (EFA) was conducted using the software Factor 12.03.02 (Ferrando & Lorenzo-Seva, 2017). Prior to this, the feasibility of conducting the exploratory factor analysis (EFA) was assessed using two standard procedures. First, Bartlett's test of sphericity was applied to verify that the correlation matrix was not an identity matrix, ensuring that the items were sufficiently interrelated to justify factor extraction. Second, sampling adequacy was evaluated using the Kaiser-Meyer-Olkin (KMO) index, both globally and at the item level through the Measure of Sampling Adequacy (MSA). The KMO provides a general indication of whether the partial correlations among items are sufficiently small, which is desirable for obtaining a reliable factor solution. The individual MSA quantifies, for each item, the proportion of shared variance (estimated through simple correlations with other items) relative to item-specific or residual variance (estimated through partial correlations) (Kaiser & Rice, 1974).

Following the criteria proposed by Lorenzo-Seva and Ferrando (2021), items whose 95% confidence intervals for the MSA did not exceed .50 were discarded. The presence of an item below this threshold indicates that it does not share sufficient variance with the rest of the scale, suggesting that it may behave erratically or measure a different construct. Retaining such items can compromise the quality and interpretability of the factor structure; therefore, their exclusion contributes to ensuring that the extracted factors are both meaningful and psychometrically sound.

Subsequently, the EFA was performed based on the tetrachoric correlation matrix, employing the robust unweighted least squares estimator. The number of factors to retain was determined using the optimal implementation of Horn's parallel analysis (Timmerman & Lorenzo-Seva, 2011). This method refines the traditional approach by generating reference eigenvalues under a randomization model through minimum rank factor analysis rather than principal component analysis, offering a more precise criterion for dimensionality assessment, particularly suitable for dichotomous data. Finally, to determine whether the data fit sufficiently well to a unidimensional solution, the traditional fit indices were obtained in their robust versions designed for categorical data: Root Mean Square Error of Approximation (RMSEA), Root Mean Square of Residuals (RMSR), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Goodness of Fit Index (GFI). Although these indices were computed using a robust procedure, their interpretation followed conventional cut-off values commonly adopted in psychometric applications: RMSEA $\leq$ .05, RMSR $\leq$ .08, CFI, TLI and GFI $\geq$ .90 (Lorenzo-Seva & Ferrando, 2023). Verifying the unidimensionality of the latent trait is a fundamental requirement for applying IRT logistic models.

There is no agreement on which dichotomous model is best for applying to the RMET items (Black, 2018; Preti et al., 2017). Therefore, we proceeded to compare the fit of the Rasch

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

(1960) model and the two-parameter (2PL) and three-parameter (3PL) logistic models (Birnbaum, 1968). The models were estimated using IRTPRO software (Cai et al., 2011). A common feature of these three models is that they are based on a logistic function, which estimates the probability of a subject responding correctly to an item given a trait value θ. The Rasch model calculates this probability using a single difficulty parameter (*bi*), representing the level of the trait required for a 0.5 probability of success. The 2PL model adds a second parameter (*ai*), which measures the item's ability to distinguish between low and high levels of the trait at its difficulty level (*bi*). Finally, the 3PL model incorporates a pseudo-guessing parameter (*ci*), which captures the likelihood of guessing correctly, even for individuals with very low levels of the latent trait.

To evaluate and compare model fit, global model fit was assessed using traditionally recommended indices (De Ayala, 2022): −2 log likelihood (−2LL), Akaike's Information Criterion (AIC), and Bayesian Information Criterion (BIC). In all cases, lower values indicate better fit. Additionally, the Likelihood Ratio Test (LRT) was conducted to assess significant differences in log-likelihood between hierarchical nested models (i.e., comparisons between 1PLM and 2PLM, as well as 2PLM and 3PLM). The test statistic, derived from the difference in −2LL values between models, follows a χ^2 distribution with degrees of freedom equal to the difference in free parameters. The models were further compared by analyzing local item independence with the χ^2 LD statistic (Chen & Thissen, 1997) and item-specific fit using the S-χ^2 statistic (Kang & Chen, 2011).

The significance of the M2 statistic as an indicator of model fit was also analyzed. This statistic, proposed by Maydeu-Olivares (2013), serves as an alternative to the conventional chi-square test for assessing the global fit of models with categorical data, such as IRT models. However, due to the excessive sensitivity of M2 to minor model-data misfit, the RMSEA was also calculated as a complementary measure. The RMSEA obtained from M2 is based on a covariance structure adjusted for dichotomous data rather than continuous covariance matrices. Consistent with the criteria used in structural equation modeling, an RMSEA ≤ .05 also indicates a satisfactory fit (Maydeu-Olivares et al., 2011).

After selecting the most suitable IRT model, item parameters were estimated using Marginal Maximum Likelihood (MML). Subsequently, item characteristic curves and the test information function were plotted, and the marginal reliability coefficient was obtained. Additionally, Cronbach's alpha and ordinal omega coefficients, along with their confidence intervals, were calculated using the *scaleStructure* function from the '*ufs*' package in R.

RESULTS

Assumption of Unidimensionality

No participant was excluded and no missed cases were reported because the protocols were completed using a virtual platform that reminded participants if any items were left unanswered.

Preliminary analyses of the items revealed the need to discard a substantial number of items due to their inadequate performance for modeling with IRT. While Bartlett's test of sphericity (χ^2 = 5689.6, *df* = 666, $p < .0001$) was statistically significant, the obtained KMO value of 0.49 was considered insufficient for conducting an exploratory factor analysis. Therefore, a total of 23 out of the 36 items (1, 2, 3, 4, 7, 8, 10, 11, 13, 16, 17, 19, 20, 23, 24, 25, 27, 28, 29, 30, 33, 34, and 35) were eliminated because the limits of the 95% confidence intervals of the

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

respective MSAs were below the .50 threshold established by Kaiser (Lorenzo-Seva & Ferrando, 2021). Appendix A contains a supplementary table with the MSA values for the 36 RMET items.

After removal of the items with low MSA values, an improvement in the estimators supporting the feasibility of the EFA was observed (KMO = .72, Bartlett's test significant, $\chi^2(78) = 1313.6$, $p < .0001$). All retained items showed MSA values $\geq .620$ (Table 1). The inter-item correlations in the original version had a mean of .058 (*SD* = .089), whereas the reduced 13-items version showed a higher mean correlation of .138 (*SD* = .091), suggesting greater internal consistency among the retained items. The mean RMET score for the full test was 25.67 (*SD* = 3.743) and for the shortened version it was 9.80 (*SD* = 2.088).

The optimal implementation of Horn's parallel analysis suggested that the 13 retained items correspond to a unidimensional structure. However, the dominant factor accounts for only 21.2% of the variance. The factor loadings were relatively low (ranging from .231 to .538), indicating that the contribution of each item is limited, and the communalities followed a similar pattern (.051 to .289), with some items capturing slightly more variance from the extracted factor (Table 1). Finally, the following fit indices were obtained: RMSEA = .043, RMSR = .073, CFI = .940, TLI = .928 and GFI = .951, complying with the conventional cut-off values commonly adopted in psychometric applications (Lorenzo-Seva & Ferrando, 2023). In sum, the results show a reasonably good fit, suggesting that the unidimensional model is viable, but with certain limitations in the variance explained and the homogeneity of the items.

Table 1. MSAs, factor loadings, and communalities from the one-factor EFA model for the 13 retained RMET items.

	MSA, IC 95%	Factor loading	Communality
Item 5	.620, [.506–.657]	.231	.053
Item 6	.625, [.521–.657]	.271	.074
Item 9	.684, [.565–.702]	.435	.189
Item 12	.647, [.538–.685]	.511	.261
Item 14	.676, [.513–.698]	.292	.085
Item 15	.697, [.559–.711]	.475	.226
Item 18	.677, [.548–.699]	.452	.204
Item 21	.685, [.506–.695]	.317	.101
Item 22	.708, [.586–.725]	.405	.164
Item 26	.712, [.583–.730]	.538	.289
Item 31	.620, [.509–.652]	.234	.055
Item 32	.652, [.503–.669]	.225	.051
Item 36	.648, [.540–.682]	.442	.195

Notes. MSA = Measure of the sample adequacy

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

Model Fit Comparison

Table 2 shows the fit statistics used to compare the fit of the three logistic models. The results consistently show that the fit of the 1PLM model is poor. Additionally, global fit indices (−2LL, AIC, BIC, M2, and RMSEA) displayed the highest values, suggesting that it is not possible to accept a constant discrimination parameter for all items (Embretson & Reise, 2013). Furthermore, the assumption of local independence was violated (χ^2LD $\geq$ 10), and the fit for four of the items was rejected, as indicated by significant S-χ^2 statistics at a 5% level.

The comparison between the 2PLM and 3PLM showed that the 2PLM was favored by lower −2LL, AIC, and BIC values. However, M2 revealed a significant reduction, indicating a better global fit for the 3PLM, suggesting that this model captures additional aspects of the item's behavior. Despite this, the RMSEA remains consistent between the 2PLM and 3PLM, suggesting that the 2PLM is sufficient for a satisfactory fit without the inclusion of parameter *c*. The verification of the local independence assumption and the fit of all items with the S-χ^2 further support the viability of both models.

The model fit comparisons using the LTR confirmed the patterns observed in the global fit indices. The comparison between the 1PLM and 2PLM showed a significant improvement in model fit (χ^2 = 101.3, *gl* = 13, $p < .001$), supporting the inclusion of item-specific discrimination parameters. In contrast, the comparison between the 2PLM and 3PLM did not yield a significant difference (χ^2 = 0.22, *gl* = 13, *n.s.*), indicating that the addition of the guessing parameter did not substantially enhance model fit.

Overall, while both the 2PLM and 3PLM models provided an acceptable fit, the 2PLM provided a better balance between fit and parsimony. The 3PLM generally provides a better fit to the data because it includes more parameters (discrimination, difficulty, and guessing). However, this comes at the cost of increased complexity and potential overfitting. On the other hand, the 2PLM is more parsimonious because it uses fewer parameters (only discrimination and difficulty), making it simpler and less prone to estimation instability issues. Given this trade-off, the 2PLM is preferable as it achieves a good fit without introducing unnecessary complexity (Embretson & Reise, 2013; Toland, 2014).

Table 2. Estimated fit indices for the Rasch, two-parameter, and three-parameter models.

	1PLM	2PLM	3PLM
−2LL	12349.61	12248.30	12248.52
AIC	12375.61	12300.30	12326.52
BIC	12438.03	12425.13	12513.77
M2	214.31	121.38	97.18
Gl	78	65	52
P	0.0001	0.0001	0.0001
RMSEA	0.04	0.03	0.03
χ^2 LD max	21.2	5.5	6.8

Note. −2LL = log likelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; M2 = M2 statistic; df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; 1PLM = Rasch Model; 2PLM = Two-parameters logistic model; 3PLM = Three-parameters logistic model.

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

Parameter Estimation and Fit of the 2PLM

A total of 24 cycles were required in the item parameter estimation process to achieve a convergence criterion of .00001. Table 3 shows the parameters of the 13 retained items after refinement, along with their respective estimation errors. Figure 1 shows the 13 characteristic curves of the items. Moreover, the S-χ^2 values with $p > .05$ are included, confirming the model's fit to the data. Difficulty parameters showed a mean of −2.07 (*SD* = 1.02). All 13 items were placed at middle-low levels on the trait scale, and even the most difficult items (items 9 and 12) did not exceed a $\theta = 0$. Furthermore, α parameters ranged from .39 to 1.22 with a mean of .78 (*SD* = .29). According to the criteria defined by Baker and Kim (2017), 46% of the items were classified as having low discrimination ($.35 \leq a \leq .64$), while 54% showed moderate discrimination ($.65 \leq a \leq 1.34$).

Figure 1 shows the item characteristic curves for the 13 retained RMET items, estimated using the 2PL model. Each curve represents the probability of selecting the correct response as a function of the latent trait level (θ). This allows for a visual inspection of each item's psychometric quality and informational value within the test. The slope of each curve reflects the item's discrimination parameter (a), that is, the degree to which the probability of a correct response changes as θ increases. This parameter indicates how well the item differentiates among individuals with varying levels of the trait. In turn, the curve's location along the x-axis corresponds to the item's difficulty parameter (b), which represents the trait level required to achieve a .50 probability of answering correctly.

Items with low discrimination (e.g., items 5, 31, and 32) exhibit curves with shallow slopes, indicating limited capacity to distinguish between individuals at different levels of the latent trait. In addition, their curves are shifted toward very low values of θ ($b < -2.5$), implying a

Table 3. Item parameters and S-χ^2 for the 13 retained items.

	α	β	χ^2	*df*	*p*
Item 5	0.40 (0.11)	−2.37 (0.63)	7.20	8	0.5167
Item 6	0.44 (0.11)	−1.70 (0.41)	9.30	8	0.3197
Item 9	0.80 (0.13)	−0.60 (0.12)	7.60	7	0.3711
Item 12	1.02 (0.15)	−0.60 (0.10)	8.80	7	0.2694
Item 14	0.58 (0.16)	−3.75 (0.92)	7.86	8	0.4489
Item 15	0.99 (0.18)	−2.16 (0.30)	4.56	8	0.8038
Item 18	0.85 (0.15)	−2.03 (0.30)	3.45	8	0.9033
Item 21	0.60 (0.14)	−3.11 (0.66)	4.26	8	0.8338
Item 22	0.77 (0.13)	−1.01 (0.17)	8.63	8	0.3759
Item 26	1.08 (0.17)	−1.56 (0.19)	3.35	8	0.9107
Item 31	0.36 (0.11)	−2.94 (0.88)	7.31	8	0.5050
Item 32	0.39 (0.12)	−3.42 (0.99)	2.51	8	0.9613
Item 36	0.91 (0.15)	−1.66 (0.23)	5.52	8	0.7017

Note. α = discrimination parameter; β = difficulty parameter; χ^2 = Chi-square; df = degrees of freedom.

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

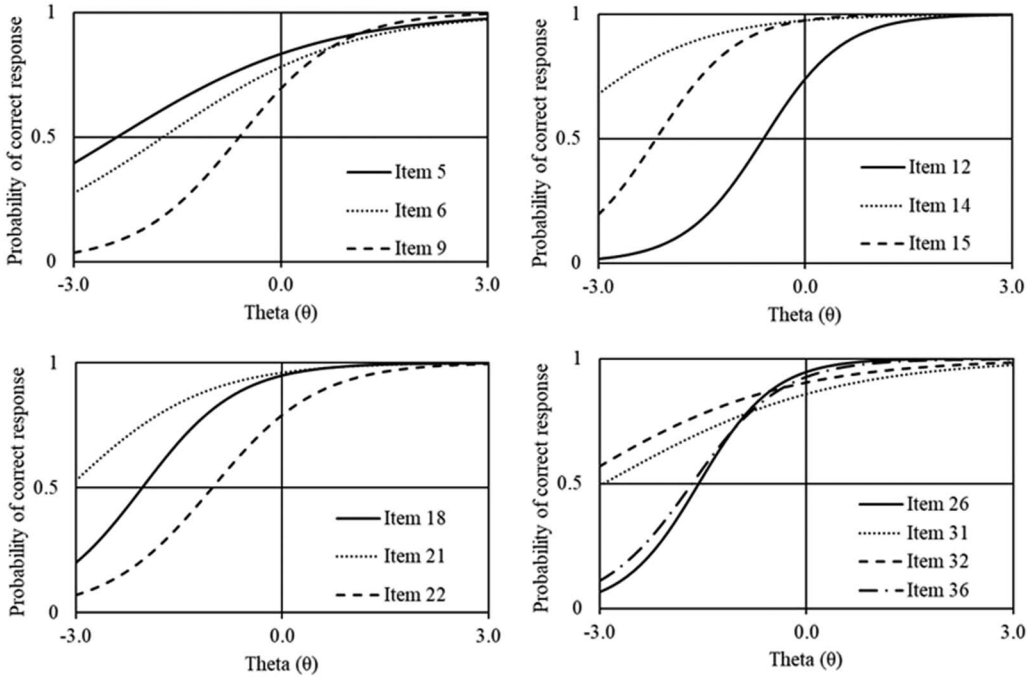


Figure 1. Item characteristic curves from the two-parameters logistic model for the 13 retained RMET items. *Note.* The retained items are grouped in the plots based on their original item numbers in the RMET. This ordering was chosen solely to enhance visual clarity and does not reflect any empirical or theoretical structure.

high probability of correct responses even at low trait levels. As a result, these items are excessively easy and provide little useful information for trait estimation.

In contrast, items with moderate discrimination and difficulty parameters between −2.0 and −0.6 (such as items 12, 15, 26, and 36) show steeper curves centered around a region of the continuum where the test achieves optimal precision—namely, the lower-middle levels of the trait. In these cases, the probability of a correct response increases rapidly as θ rises, allowing for more effective differentiation among individuals with low to moderately low trait levels.

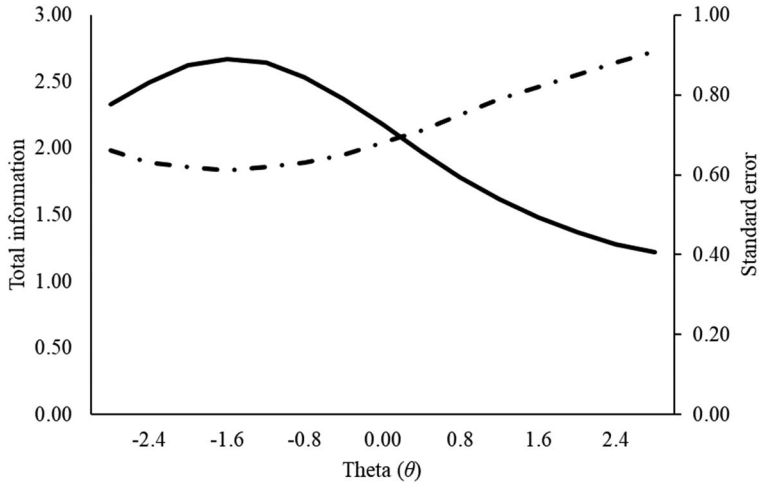


Figure 2. Test information and standard error functions for the 2PL model estimated from the 13 retained RMET items.

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

Although this subset of items has acceptable psychometric properties, their small number limits the overall precision of trait measurement.

Reliability Analysis

Results showed that internal consistency estimates obtained under classical test theory (CTT) were deficient. The ordinal alpha coefficient was .41, 95% CI [.35, .47], while McDonald's omega was .11, 95% CI [.06, .17]. Furthermore, the marginal reliability coefficient obtained under IRT showed a value of .52. The test information function (TIF) reached a maximum value of 2.67 for $\theta = -1.60$ (Figure 2). Moreover, for this trait level, the standard error function reached its lowest value of .61 (equivalent to a classical reliability of .63).

DISCUSSION

Even though the RMET was developed almost 30 years ago, it has maintained its relevance through widespread use and adaptations across various cultural contexts. However, discussions regarding its psychometric properties are relatively recent, and scientific evidence suggests that the test has certain limitations that place it below the recommended standards for psychological research (Higgins, Kaplan, et al., 2024).

The present study employed IRT to examine the psychometric properties of the RMET items in its Argentinian version. Moreover, as IRT offers a more detailed and rigorous approach compared to classical test theory, it is particularly suitable for assessing performance tasks like the RMET (Reise & Revicki, 2015; Thomas, 2019).

Initially, an EFA was conducted to assess the construct's dimensionality, which is a procedure required for the application of the IRT logistic models. Of the few studies that have explored the internal structure of the RMET, most have examined a unidimensional model (Đorđević et al., 2017; Kotrla Topić & Perković Kovačević, 2019; Vellante et al., 2013), which aligns with the technical assumption of the test, as a single total score is commonly used. However, these studies do not provide clear support for a unidimensional model, and multidimensional structures of the construct have also been explored (Kittel et al., 2022). For instance, the Korean version of the RMET presents a three-factor structure based on the emotional valences, consisting of positive, negative, and neutral (Lee et al., 2020).

In this study, a preliminary item analysis conducted prior to the EFA revealed the need to refine the test, leading to the elimination of a considerable number of items (24 items). The criterion used, based on the MSA value, aligns with current guidelines that recommend removing "noisy items"—which may behave nearly randomly—before performing the EFA (Lorenzo-Seva & Ferrando, 2021). These findings align with several studies that decided to remove items from the RMET due to their poor psychometric properties (Black, 2018; Đorđević et al., 2017; Kotrla Topić & Perković Kovačević, 2019; López & Richaud de Minzi, 2020; Olderbak et al., 2015).

Correspondingly, even though the 13 items retained displayed a unidimensional structure, the dominant factor accounts for only 21.2% of the variance, with relatively low factor loadings ranging from .23 to .54. This aligns with findings from other studies, which suggest that the dimensionality of the construct exhibits weak and questionable evidence (Higgins, Kaplan, et al., 2024; Olderbak et al., 2015).

Moreover, these weak psychometric properties are also observed under the IRT analysis. All 13 items presented medium-low difficulty parameters ($\beta \leq -0.60$) and discrimination values (α) indicating only a moderate capacity ($0.36 \leq \alpha \leq 1.08$). Similarly, Black (2018) applied the

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

2PLM to a 20 items model, most of which showed values below the trait mean ($\beta \leq 0.01$) and displayed slightly higher discrimination values ($0.70 \leq a \leq 1.46$). Alternatively, Preti et al. (2017) reported more encouraging results, successfully fitting the Rasch Model for 32 out of the 36 items. However, these results were based on a relatively small sample size ($n = 200$), which could affect the stability and generalizability of these findings.

Regarding the conceptual characteristics of these 13 items, we did not detect an observable pattern that would suggest that they better capture ToM compared to the excluded items. Of the 13 items, 7 are photographs of women and 6 of men; 8 are photographs of direct gaze and 5 of lateral gaze. Ten items (5, 6, 12, 14, 15, 18, 26, 31, 32 and 36) of the 13 preserved in this abbreviated Argentine version correspond to the items preserved in the version made López and Richaud de Minzi (2020). Only 6 items (5, 9, 15, 18, 32 and 36) of the 13 preserved in this abbreviated Argentine version correspond to the items preserved in the version made by Black (2018). In this Argentine version, 5 of the 6 elements that were eliminated from the short Brazilian version were removed (Miguel et al., 2017). However, it should be noted that these 13 retained items in the Argentine version include, among their response options, concepts that are infrequently used in the Argentine context, such as: "abochornado", "implorante" y "acusante", which possibly reveals deficiencies in the test translation, which makes comparison with versions in other languages difficult.

Considering these results and the limited prior evidence on the robust RMET's factor structure (Higgins, Kaplan, et al., 2024; Higgins, Savalei, et al., 2024), it is important to determine whether the issue lies in the theoretical nature of the construct or its operationalization.

Regarding the theoretical aspect of the construct, the discrepancies in dimensionality and limited discrimination capacity could partly be explained by the fact that, although many idiosyncratic definitions of ToM have been developed, there is no clear model of its structure. This lack of clarity often leads to ToM being considered a construct similar to empathy, as both involve inferring the mental state of others. Moreover, it has been suggested that there are different levels or degrees of ToM, and even different tests were developed to assess its components (Kittel et al., 2022). Hence, it might not be appropriate to categorize the RMET as a general ToM test, but rather as a specific task for decoding mental states through individuals' gaze expressions (Atenas et al., 2019).

In fact, the RMET's design is similar to emotion perception tests; however, unlike those tests, it focuses solely on the eyes (Kittel et al., 2022). At a conceptual level, a distinction could be made between low-level perceptual processing and the decoding of emotional states that involve emotion recognition, and the high-level cognitive process of inferring mental states associated with ToM (Stafford et al., 2023). Nevertheless, studies show high correlations between the RMET and emotion perception tests, with a shared variance of 23%, which is higher than the correlation with other ToM tests such as the Faux Pas Test (Kittel et al., 2022).

Furthermore, the theoretical and methodological basis of emotion and mental state assessments is currently being questioned. Several studies (including meta-analyses) do not support the hypothesis that emotions or mental states present specific and synchronized brain patterns with measurable changes on the face. Moreover, research using facial electromyography does not support the existence of unique facial expressions for each emotion, as people display varied facial movements for emotions such as happiness or anger (Feldman-Barrett & Westlin, 2021; Šimić et al., 2021). From this new perspective in neuroscience, it may not be possible to recognize emotions or mental states, since their perception is influenced by the context. Therefore, they cannot be effectively assessed through facial expression analysis alone but

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

require a multimodal measurement framework (Feldman-Barrett & Westlin, 2021), which should begin to be considered not only in research but especially in clinical practice.

Regarding the operationalization of the construct, both the development of the assessment and the criteria used to determine the validity of the correct answers are questionable, especially in regard to whether a "correct" answer truly exists. Baron-Cohen et al. (2001) selected photographs for the RMET from magazines without knowing what emotions the models were attempting to express, and later chose four terms related to mental states for each of the items considered as a candidate for inclusion in the RMET. The test was validated in a sample of 103 undergraduate students from Cambridge University and 122 participants from the general population in the UK. Subsequently, two validation criteria were applied to retain an item: a) at least 50% of participants must choose the option as 'correct'; b) no more than 25% of participants should select the same alternative as 'incorrect'. However, these criteria are not consistently met across the different populations studied in the RMET research, which suggests that cultural differences may contribute to the differences found across studies (Higgins, Savalei, et al., 2024). On the other hand, the need for a 50% minimum consensus for the correct answer may have led to the inclusion of items with medium-low difficulty, consistent with the estimated values for the β parameters in 2PLM. Furthermore, in certain cultures, higher difficulty parameters might result in lower consensus among participants, potentially reducing the items' ability to discriminate effectively. As a result, responses to the items can barely be assessed using a true response standard, indicating limitations in the design of the instrument (Kittel et al., 2022).

Together with the absence of consistent evidence concerning its internal structure, this raises doubts about whether the RMET accurately measures ToM. This is further supported by the global reliability indices, derived from both classical test theory and IRT, which were found to be inadequate; results that are consistent with previous research (Higgins, Kaplan, et al., 2024; Olderbak et al., 2015). Moreover, even in a large-scale RMET study by Kittel et al. (2022), it is concluded that the acceptable internal consistency reported in some studies appears to be a result of the test's length.

Conversely, the item information scores from IRT revealed that the 13 retained items only capture information for the lower to mid-levels of the trait, while precision drops significantly at higher levels due to the lack of items positioned at those levels. Furthermore, Black (2018) similarly suggest that the RMET is a test that fails to distinguish between individuals with normal or high abilities, thereby limiting its capacity to identify individual differences in neurotypical populations. According to the author, although the RMET can be helpful in identifying mentalization deficits typically seen in individuals with high-functioning autism, it is less effective at distinguishing higher levels of ToM due to the absence of sufficiently difficult items for this purpose. Therefore, the range restriction caused by the ceiling effect of the scores highlights the need for caution when making inferences from studies that have used this test as a measure of social cognition in non-clinical populations (Black, 2018).

It is important to consider that the original design of the task in 1997 aimed to measure social cognitive skills in children and adults with autism, with an emphasis on its ability to differentiate among individuals with low levels of ToM. However, in 2001, the test was revised (Baron-Cohen et al., 2001), increasing the number of items and response options with greater difficulty, to minimize the potential for ceiling effects in typical performance. As a result, the test's applicability was extended, and it was concluded that the RMET could also detect differences in neurotypical adults' ability to infer mental states from minimal cues such as the eyes' expression (Yıldırım et al., 2011). Nevertheless, the results of the present study indicate

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

that this test remains an unreliable assessment tool for the Argentine neurotypical population, in addition to the limitations of the test that have already been identified in earlier studies (for more details, see Kittel et al., 2022).

In conclusion, the results of this study, consistent with previous research, are discouraging and do not support the psychometric properties of the RMET. These findings raise questions about the validity of the measure and the conclusions drawn from studies using it in the extensive scientific literature. Furthermore, these results are part of the growing evidence pointing to the so-called "measurement crisis" in psychological research, according to which researchers do not rigorously and systematically examine whether psychological measures truly assess what they are intended to measure (Flake & Fried, 2020). Therefore, this study aimed to contribute to increasing awareness of the critical need to thoroughly evaluate the validity and reliability of the RMET as an assessment tool, alongside the theoretical construct it operationalizes, prior to its continued use in research. Moreover, given the clinical, diagnostic, and prognostic significance attributed to the scores of this instrument for evaluating Theory of Mind or mentalization (Higgins, Kaplan, et al., 2024), it is frequently used in clinical research where individuals with clinical conditions, including autism spectrum disorder and schizophrenia, have been found to score lower on average than non-clinical groups (Alvarez et al., 2023; Deng et al., 2024; Johnson et al., 2022; Peñuelas-Calvo et al., 2019). Consequently, it is recommended that researchers and clinicians approach the administration and interpretation of RMET results with caution (Kittel et al., 2022).

Among the key limitations of this research is the data collection process, which was conducted online in a self-administered manner, without supervision during the test administration. This method, though widely used to achieve large sample sizes (Black, 2018; Kittel et al., 2022; Lee et al., 2020; López & Richaud de Minzi, 2020), might affect the quality of the data. Specifically, a recent Argentine study (Carrillo, 2021) has shown that there are statistically significant differences between the administration of the RMET in virtual and pencil-and-paper formats, with higher scores being obtained in the virtual format. Furthermore, the analyses in this study focused solely on assessing reliability and gathering validity evidence based on the internal structure, ignoring the relationship with external variables and the test's diagnostic value by contrasting it with clinical population groups. Moreover, the sample was predominantly composed of individuals with university studies, either in progress or completed, and from urban areas. Therefore, although IRT allows the estimation of population-invariant parameters when its assumptions are met (Embretson & Reise, 2013), future studies should focus on including more diverse demographic groups to enhance the generalizability of the findings.

AUTHOR CONTRIBUTIONS

R.G.: Data curation; Investigation; Resources; Writing – original draft. J.I.L.: Data curation; Investigation; Methodology; Writing review & editing. F.J.P.A.: Formal analysis; Methodology; Writing review & editing.

FUNDING INFORMATION

This work was supported by a doctoral grant awarded to the first author by the "Consejo Nacional de Investigaciones Científicas y Técnicas" (CONICET), Argentina.

DATA AVAILABILITY STATEMENT

Data used in the present study are not publicly available. For inquiries about data access please contact the first author.

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

REFERENCES

Alvarez, R., Velthorst, E., Pinkham, A., Ludwig, K. A., Alamansa, J., Gaigg, S. B., ... Fett, A.-K. (2023). Reading the mind in the eyes and cognitive ability in schizophrenia- and autism spectrum disorders. *Psychological Medicine*, *53*(16), 7913–7922. https://doi.org/10.1017/S0033291723002052, PubMed: 37522512

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (DSM-5®). American Psychiatric Pub. https://doi.org/10.1176/appi.books.9780890425596

Atenas, T. L., Ciampi Diaz, E., Venegas Bustos, J., Uribe San Martín, R., & Cárcamo Rodríguez, C. (2019). Cognición social: Conceptos y bases neurales. *Revista Chilena de Neuropsiquiatría*, *57*(4), 365–376. https://doi.org/10.4067/S0717-92272019000400365

Ato, M., López-García, J. J., & Benavente, A. (2013). Un sistema de clasificación de los diseños de investigación en psicología. *Anales de Psicología*, *29*(3), 1038–1059. https://doi.org/10.6018/analesps.29.3.178511

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*. Springer. https://doi.org/10.1007/978-3-319-54205-8

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.2044-835X.1995.tb00687.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. https://doi.org/10.1017/S0021963001006643, PubMed: 11280420

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of mental test scores*. Reading, Mass: Addison-Wesley.

Black, J. E. (2018). An IRT analysis of the reading mind in the eyes test. *Journal of Personality Assessment*, *101*(4), 425–433. https://doi.org/10.1080/00223891.2018.1447946, PubMed: 29611711

Cai, L., Thissen, D., & du Toit, S. (2011). *IRTPRO user's guide*. Lincolnwood, IL: Scientific Software International.

Carrillo, C. (2021). Estudio de comparación entre el formato virtual y lápiz-papel en la evaluación de la teoría de la mente. In *XIII Congreso Internacional de Investigación y Práctica Profesional en Psicología*. XXVIII Jornadas de Investigación. XVII Encuentro de Investigadores en Psicología del MERCOSUR. III Encuentro de Investigación de Terapia Ocupacional. III Encuentro de Musicoterapia. Facultad de Psicología-Universidad de Buenos Aires.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Chung, D., & Yoon, W. (2019). P14.51 Brain mapping for mentalizing in GBM patient. *Neuro-Oncology*, *21*(Suppl. 3), iii78–iii79. https://doi.org/10.1093/neuonc/noz126.286

Cohen, R. F., Berardi, A. M., Le Vaou, P., & Kahn, J.-P. (2019). Validation and psychometric properties of the "Reading the Mind in the Eyes Test". Preprint version.

de Ayala, R. J. (2022). *The theory and practice of item response theory*. Guilford Press.

Deng, F., Bueber, M. A., Cao, Y., Tang, J., Bai, X., Cho, Y., ... Phillips, M. R. (2024). Assessing social cognition in patients with schizophrenia and healthy controls using the reading the mind in the eyes test (RMET): A systematic review and meta-regression. *Psychological Medicine*, *54*(5), 847–873. https://doi.org/10.1017/S0033291723003501, PubMed: 38173096

Đorđević, J., Živanović, M., Pavlović, A., Mihajlović, G., Stašević, K. I., & Pavlović, D. (2017). Psychometric evaluation and validation of the Serbian version of 'Reading the Mind in the Eyes' test. *Psihologija*, *50*(4), 483–502. https://doi.org/10.2298/PSI170504010D

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Feldman-Barrett, L., & Westlin, C. (2021). Navigating the science of emotion. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 39–84). Woodhead Publishing. https://doi.org/10.1016/B978-0-12-821124-3.00002-8

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*(2), 236–240. https://doi.org/10.7334/psicothema2016.304, PubMed: 28438248

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

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, PubMed: 38232573

Higgins, W. C., Savalei, V., Polito, V., & Ross, R. M. (2024). Reading the Mind in the Eyes Test scores demonstrate poor structural properties in nine large non-clinical samples. *PsyArXiv* Preprint version.

Johnson, B. N., Kivity, Y., Rosenstein, L. K., LeBreton, J. M., & Levy, K. N. (2022). The association between mentalizing and psychopathology: A meta-analysis of the reading the mind in the eyes task across psychiatric disorders. *Clinical Psychology: Science and Practice*, *29*(4), 423–439. https://doi.org/10.1037/cps0000105

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kang, T., & Chen, T. T. (2011). Performance of the generalized S-X2 item fit index for the graded response model. *Asia Pacific Education Review*, *12*(1), 89–96. https://doi.org/10.1007/s12564-010-9082-4

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, PubMed: 33645295

Kliemann, D., Adolphs, R., Paul, L. K., Tyszka, J. M., & Tranel, D. (2021). Reorganization of the social brain in individuals with only one intact cerebral hemisphere. *Brain Sciences*, *11*(8), 965. https://doi.org/10.3390/brainsci11080965, PubMed: 34439583

Kotrla Topić, M., & Perković Kovačević, M. (2019). Croatian adaptation of the revised reading the mind in the eyes test (RMET). *Psihologijske Teme*, *28*(2), 377–395. https://doi.org/10.31820/pt.28.2.8

Lee, H.-R., Nam, G., & Hur, J.-W. (2020). Development and validation of the Korean version of the Reading the Mind in the Eyes Test. *PLoS One*, *15*(8), e0238309. https://doi.org/10.1371/journal.pone.0238309, PubMed: 32866184

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

López, M. B., & Richaud de Minzi, M. C. (2020). Estudio de validez y consistencia interna de una versión en español del test "reading the mind in the eyes" y propuesta de una versión breve. *Cuadernos de Neuropsicologia*, *14*(1), 88–102.

López, M. B., Arán Filippetti, V., & Richaud de Minzi, M. C. (2014). Empatía: Desde la percepción automática hasta los procesos controlados. *Avances en Psicología Latinoamericana*, *328*(1), 37–51. https://doi.org/10.12804/apl32.1.2014.03

Lorenzo-Seva, U., & Ferrando, P. J. (2021). MSA: The forgotten index for identifying inappropriate items before computing exploratory item factor analysis. *Methodology*, *17*(4), 296–306. https://doi.org/10.5964/meth.7185

Lorenzo-Seva, U., & Ferrando, P. J. (2023). A simulation-based scaled test statistic for assessing model–data fit in least-squares unrestricted factor-analysis solutions. *Methodology*, *19*(2), 96–115. https://doi.org/10.5964/meth.9839

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71–101. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333–356. https://doi.org/10.1080/10705511.2011.581993

Miguel, F. K., Caramanico, R. B., Huss, E. Y., & Zuanazzi, A. C. (2017). Validity of the reading the mind in the eyes test in a Brazilian sample. *Paidéia (Ribeirão Preto)*, *27*(66), 16–23. https://doi.org/10.1590/1982-43272766201703

National Institute of Mental Health. (2016). Behavioral assessment methods for RDoC constructs. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/understanding-mental-states.shtml

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*, 1503. https://doi.org/10.3389/fpsyg.2015.01503, PubMed: 26500578

Peñuelas-Calvo, I., Sareen, A., Sevilla-Llewellyn-Jones, J., & Fernández-Berrocal, P. (2019). The "Reading the Mind in the Eyes" test in autism-spectrum disorders comparison with healthy controls: A systematic review and meta-analysis. *Journal of Autism and Developmental Disorders*, *49*(3), 1048–1061. https://doi.org/10.1007/s10803-018-3814-4, PubMed: 30406435

Preti, A., Vellante, M., & Petretto, D. R. (2017). The psychometric properties of the "Reading the Mind in the Eyes" Test: An item response theory (IRT) analysis. *Cognitive Neuropsychiatry*, *22*(3), 233–253. https://doi.org/10.1080/13546805.2017.1300091, PubMed: 28288549

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling applications to typical performance assessment*. Routledge.

Roman, F. N., Rojas, G., Roman, N. R., Iturry, M., Blanco, R., Leis, A., ... Allegri, R. F. (2012). Baremos del Test de la Mirada en español en adultos normales de Buenos Aires. *Neuropsicología Latinoamericana*, *4*(3), 1–5. https://doi.org/10.5579/rnl.2012.0108

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., ... Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, *147*(3), 293–327. https://doi.org/10.1037/bul0000303, PubMed: 33151703

Šimić, G., Tkalčić, M., Vukić, V., Mulc, D., Španić, E., Šagud, M., ... Hof, P. R. (2021). Understanding emotions: Origins and roles of the amygdala. *Biomolecules*, *11*(6), 823. https://doi.org/10.3390/biom11060823, PubMed: 34072960

Stafford, O., Gleeson, C., Egan, C., Tunney, C., Rooney, B., O'Keeffe, F., ... Burke, T. (2023). A 20-year systematic review of the 'Reading the Mind in the Eyes' test across neurodegenerative conditions. *Brain Sciences*, *13*(9), 1268. https://doi.org/10.3390/brainsci13091268, PubMed: 37759869

Thomas, M. L. (2019). Advances in applications of item response theory to clinical assessment. *Psychological Assessment*, *31*(12), 1442–1455. https://doi.org/10.1037/pas0000597, PubMed: 30869966

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, *16*(2), 209–220. https://doi.org/10.1037/a0023353, PubMed: 21500916

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *Journal of Early Adolescence*, *34*(1), 120–151. https://doi.org/10.1177/0272431613511332

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, PubMed: 23106125

Yıldırım, E. A., Kaşar, M., Güdük, M., Ateş, E., Küçükparlak, I., & Özalmete, E. O. (2011). Investigation of the reliability of the "reading the mind in the eyes test" in a Turkish population. *Turkish Journal of Psychiatry*, *22*(3), 177–186. PubMed: 21870307

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

APPENDIX A

Table A1. Measures of Sampling Adequacy (MSAs) for the 37 RMET items.

	MSA, IC 95%
Item 1	.607, [.490–.641]
Item 2	.567, [.459–.618]
Item 3	.553, [.453–.608]
Item 4	.570, [.481–.611]
Item 5*	.620, [.506–.657]
Item 6*	.625, [.521–.657]
Item 7	.535, [.443–.583]
Item 8	.618, [.498–.649]
Item 9*	.684, [.565–.702]
Item 10	.606, [.471–.617]
Item 11	.584, [.486–.619]
Item 12*	.647, [.538–.685]
Item 13	.536, [.435–.592]
Item 14*	.676, [.513–.698]
Item 15*	.697, [.559–.711]
Item 16	.574, [.463–.612]
Item 17	.573, [.476–.625]
Item 18*	.677, [.548–.699]
Item 19	.607, [.481–.647]
Item 20	.649, [.498–.678]
Item 21*	.685, [.506–.695]
Item 22*	.708, [.586–.725]
Item 23	.620, [.479–.655]
Item 24	.554, [.445–.613]
Item 25	.604, [.484–.639]
Item 26*	.712, [.583–.730]
Item 27	.592, [.480–.635]
Item 28	.589, [.469–.643]
Item 29	.578, [.476–.630]

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026

Table A1. (*continued*)

	MSA, IC 95%
Item 30	.570, [.465–.636]
Item 31*	.620, [.509–.652]
Item 32*	.652, [.503–.669]
Item 33	.561, [.461–.613]
Item 34	.566, [.473–.629]
Item 35	.600, [.480–.630]
Item 36*	.648, [.540–.682]

Note. Asterisks indicate items retained for further analyses based on MSA criteria.

Downloaded from http://direct.mit.edu/opmi/article-pdf/doi/10.1162/opmi.a.11/2547357/opmi.a.11.pdf by guest on 06 January 2026