

Article

Psychometric Evaluation and Normative Framework for the Movie for the Assessment of Social Cognition in Spanish-Speaking Adults

Aina Llompart-Casasnovas¹, Adoración Castro^{1,2}, Margalida Gili^{1,2}, Guillem Navarra-Ventura^{1,2}, Roberto Rodríguez-Jiménez^{3,4}, Miguel A. Ortega⁵, Leticia Boada^{4,6}, Mercè Jodar^{4,7}, Narcís Cardoner^{4,8}, Albert Sesé^{1,2}, Miquel Roca^{1,2} and Guillermo Lahera^{4,5}

¹ University of the Balearic Islands (Spain)

² Health Research Institute of the Balearic Islands (Spain)

³ Universidad Complutense de Madrid (Spain)

⁴ CIBER de Salud Mental (Spain)

⁵ Universidad de Alcalá (Spain)

⁶ Hospital Gregorio Marañón (Spain)

⁷ Hospital Universitari Parc Taulí (Spain)

⁸ Hospital de la Sant Creu i Sant Pau (Spain)

ARTICLE INFO

Received: 03/10/2025

Accepted: 24/03/2026

Keywords:

Theory of mind
Mentalization
Factor analysis
Internal consistency
Percentiles

ABSTRACT

Background: The Movie for the Assessment of Social Cognition-Spanish (MASC-SP) assesses general mentalization performance and associated error patterns: over-mentalization, under-mentalization, and non-mentalization. This study aims to provide the first large-scale psychometric evaluation and normative framework for MASC-SP scores in Spanish-speaking adults. **Method:** 2,582 healthy adults completed the MASC-SP via a secure online platform. The psychometric evaluation included factor analysis, measurement invariance, internal consistency, and floor/ceiling effects. Normative data were stratified by demographic variables and are presented as empirical percentiles and z-scores. **Results:** MASC-SP scores showed a predominantly unifactorial structure, with internal consistency ranging from questionable to good. A floor effect was observed in the non-mentalization error score. Female participants and those with a university education scored higher and made fewer errors than men and those without a university education. Age-related results were heterogeneous. **Conclusions:** The results support interpreting the MASC-SP total score as a robust quantitative indicator of general mentalization performance and the error scores as secondary qualitative indicators of different error patterns. Normative data stratified by sex and educational level provide a useful reference for research and clinical settings. However, norms stratified by age groups should be applied with caution due to their questionable clinical and developmental relevance.

Evaluación Psicométrica y Marco Normativo para la *Movie for the Assessment of Social Cognition* en Adultos Castellanohablantes

RESUMEN

Antecedentes: La *Movie for the Assessment of Social Cognition-Spanish* (MASC-SP) evalúa rendimiento mentalizador general y patrones de error asociados: sobre-mentalización, infra-mentalización y no-mentalización. El objetivo es proporcionar la primera evaluación psicométrica y marco normativo a gran escala para las puntuaciones de la MASC-SP en adultos castellanohablantes. **Método:** 2.582 adultos sanos completaron la MASC-SP mediante una plataforma en línea. La evaluación psicométrica incluyó análisis factorial, invarianza de medida, consistencia interna y efectos suelo/techo. Los datos normativos se estratificaron según variables demográficas y se presentan como percentiles y puntuaciones *z* empíricas. **Resultados:** Las puntuaciones mostraron una estructura predominantemente unifactorial, con una consistencia interna entre cuestionable y buena. Se observó un efecto suelo en la puntuación de error de no-mentalización. Las mujeres y los participantes con estudios universitarios obtuvieron puntuaciones más altas y cometieron menos errores que sus contrapartes. Los resultados relacionados con la edad fueron heterogéneos. **Conclusiones:** Los resultados respaldan la interpretación de la puntuación total de la MASC-SP como un indicador cuantitativo sólido del rendimiento mentalizador general y de las puntuaciones de error como indicadores cualitativos secundarios. Los datos normativos estratificados por sexo y educación proporcionan una referencia útil para contextos de investigación y clínicos. Las normas estratificadas por grupos de edad deben aplicarse con cautela debido a su cuestionable relevancia clínica y de desarrollo.

Palabras clave:

Teoría de la mente
Mentalización
Análisis factorial
Consistencia interna
Percentiles

Social cognition (SC) enables the perception, processing, and interpretation of social information, facilitates understanding of others' behaviors and intentions, and promotes effective functioning in social interactions (Arioli et al., 2018). It is a multidimensional and hierarchically organized construct that encompasses several interrelated domains, of which the most studied are emotional processing and theory of mind (ToM) (Schurz et al., 2021). To date, research has focused on key subcomponents, such as emotion recognition, mental state reasoning (inferring thoughts/beliefs), and mental state decoding (inferring feelings) (Mitchell & Phillips, 2015).

Patients with mental disorders often present SC deficits (Cotter et al., 2018), which tend to be more pronounced in the more complex, higher-order subcomponents than in the more basic subcomponents (Navarra-Ventura et al., 2022; Romero-Ferreiro et al., 2022). These deficits have been documented in Asperger syndrome (Dziobek et al., 2006; Lahera et al., 2014), schizophrenia (Andrade-González et al., 2021; Catalan et al., 2018; Engelstad et al., 2019; Fretland et al., 2015; Montag et al., 2011; Sahl et al., 2023; Vaskinn et al., 2018), bipolar disorder (Aidelbaum & Goghari, 2022; Montag et al., 2010; Santos et al., 2017), depression (Seitz et al., 2022; Wolkenstein et al., 2011; Zwick & Wolkenstein, 2017), and personality disorder (Andreou et al., 2015; Newbury-Helps et al., 2017; Preißler et al., 2010; Quek et al., 2018; Somma et al., 2019).

Demographic factors also influence SC performance. Older age and lower educational level have been associated with reduced mental state reasoning and decoding, while emotion recognition appears to be less affected by these factors (Fittipaldi et al., 2024). Other studies have revealed sex-related differences in temporo-limbic and frontal brain regions (Gur, 2002), and in large-scale connectivity patterns (Ingallhalikar et al., 2014), which are often accompanied by behavioral differences. Women tend to outperform men in emotion recognition, while sex differences in mental state reasoning and decoding are generally small or non-existent (Christov-Moore et al., 2014; Navarra-Ventura et al., 2018, 2021, 2022; Vaskinn et al., 2024).

These findings underscore the multifactorial and heterogeneous nature of SC and highlight the need for assessment instruments that are psychometrically sound and supported by normative data (Pinkham et al., 2018). Traditional instruments include the Pictures of Facial Affect (POFA) for emotion recognition (Ekman & Friesen, 1976), the Strange Stories Test (SST) for mental state reasoning (Happé, 1994), and the Reading the Mind in the Eyes Test (RMET) for mental state decoding (Baron-Cohen et al., 2001). Despite their widespread use, the scores of these instruments are limited by the presence of ceiling effects, their low ecological validity, and their exclusive focus on a single subcomponent of SC (Dziobek et al., 2006; Lahera et al., 2014). In contrast, real-life social interactions require the simultaneous integration of multiple dynamic signals and cues (facial expression, gaze, gestures, body language, speech prosody, verbal content) to understand others' behaviors and intentions (Cavieres et al., 2025).

Among the instruments designed to capture this complexity, the Movie for the Assessment of Social Cognition (MASC) (Dziobek et al., 2006) stands out. The MASC integrates verbal and non-verbal signals and cues related to emotional processing and ToM in video format, and its scores provide a more ecologically valid and dynamic assessment of SC than traditional instruments. Its multiple-choice question structure allows for distinguishing between typical

and impaired ToM performance (primary outcome) and identifying specific error patterns, including over-attribution, under-attribution, and non-attribution of mental states (secondary outcomes). This scoring system allows for a more nuanced assessment of mentalization abilities and associated error patterns, thereby improving the interpretability of results (Msika et al., 2024).

To date, the MASC has been adapted into several languages, including Spanish (MASC-SP) (Lahera et al., 2014), and there is growing evidence supporting the clinical sensitivity and cross-cultural applicability of its scores (Chang et al., 2024; Fossati et al., 2018; Huang et al., 2023; Martinez et al., 2017; Müller et al., 2016; Sarig-Shmueli et al., 2023). However, these studies are limited by small sample sizes, lack of psychometric evaluation of secondary scores, and lack of normative data, which has restricted their use to research settings and their applicability in clinical practice.

This study aims to provide the first large-scale psychometric evaluation and normative framework for MASC-SP scores in Spanish-speaking adults.

Method

Participants

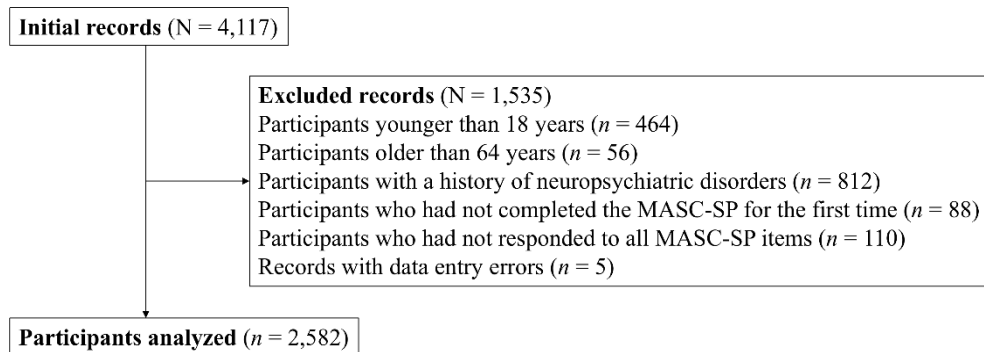
All native Spanish speakers residing in Spain were eligible to participate. Participants were recruited between June 2016 and May 2024 using non-probability methods, primarily convenience and snowball sampling, in community organizations, universities, and clinical settings through direct contact and study invitations.

The initial dataset comprised 4,117 records, which were filtered according to the following exclusion criteria: the participant was under 18 or over 64 years of age, had a history of neuropsychiatric disorders, had previously completed the MASC-SP, or had not completed all items on the MASC-SP. Five additional records were excluded due to data entry errors. After applying these criteria, 1,535 records were excluded, resulting in a final sample of 2,582 participants (Figure 1).

Instruments

The MASC-SP (Lahera et al., 2014) assesses mentalization abilities through real-life social situations presented in video format. The video shows four people (two men, two women) interacting at a dinner party, and participants are asked to infer the characters' beliefs/thoughts (mental state reasoning) and feelings (mental state decoding). The video pauses several times to present up to 45 multiple-choice questions about the characters' behaviors and intentions. Each question includes four answer options: one correct and three incorrect, allowing for the identification of different types of mentalization errors: over-mentalization (attributing a mental state when no mental explanation is required), under-mentalization (misattributing an existing mental state), and non-mentalization (total absence of mental inference, such as attributing physical causality to social situations and mental states). An example of the items and their response options can be found elsewhere (Dziobek et al., 2006; Lahera et al., 2014). A primary total score for correct answers, ranging from 0 (worst mentalization) to 45 (best mentalization), and three secondary scores for errors, ranging from 0 (no errors) to 45 (maximum errors), are calculated. The video lasts 15 minutes and

Figure 1
Flow Diagram



Note. MASC-SP, Movie for the Assessment of Social Cognition-Spanish.

participants have up to 30 seconds to answer each question, giving a total administration time of 45 minutes.

Studies examining the psychometric properties of the MASC and MASC-SP have primarily focused on the total score. In the original study (Dziobek et al., 2006), the total score effectively discriminated adults with Asperger syndrome from matched control group participants, achieving an area under the curve of .98. This performance exceeded that of the total scores of the POFA (.79), SST (.65), and RMET (.86). In the Spanish adaptation study (Lahera et al., 2014), moderate-to-strong Pearson correlations were found between the total score and the total scores of the POFA (.57), SST (.68), and RMET (.77), supporting its convergent validity. In contrast, correlations with general cognition, as measured by the Screen for Cognitive Impairment in Psychiatry, were not significant, providing evidence of discriminant validity. The total score also showed optimal internal consistency ($\alpha = .86$) and good test-retest reliability.

Procedure

The study was approved by the Ethics Committee of the University of the Balearic Islands in Mallorca (Balearic Islands), Spain (#011CER25), and was conducted in accordance with the principles of the Declaration of Helsinki (1964) and its subsequent amendments. The study was not preregistered.

All participants completed the MASC-SP via a password-protected online platform (<http://masc-cognicionsocial.es/>), supported by the Instrument Library of the Center for Biomedical Research in Mental Health Network (CIBERSAM) at the Carlos III Health Institute (ISCIII) in Madrid, Spain. Data were also collected on sex (male/female), age (in whole years), educational level (non-university/university), history of neuropsychiatric disorders (yes/no), and whether it was the first time the participant had completed the MASC-SP (yes/no). Participants who did not complete all items on the MASC-SP were excluded from the analysis due to missing data (Figure 1).

Data Analysis

Data normality was verified using histograms, skewness, kurtosis, and the Kolmogorov-Smirnov test. When score distributions deviated from normality, non-parametric analyses were used. Descriptive statistics are presented as frequencies and percentages, means (M) and standard deviations (SD), or medians (Mdn) and ranges (minimum-maximum). The effect size r is also reported. Analyses were performed using IBM SPSS Statistics 29, with statistical significance set at $p < .05$.

The psychometric properties of the MASC-SP scores were examined using Spearman correlations, exploratory factor analysis (EFA) with principal axis factoring and Oblimin rotation, and Cronbach's alpha for internal consistency. For correlations, scores were treated as continuous variables. For EFA and internal consistency, item-level responses were dichotomized, resulting in four binary indicators per item corresponding to total and error scores. We further examined the dimensional structure of the total score using EFA on the tetrachoric correlation matrix, as implemented in the R packages psych (Revelle, 2025) and GPArotation (Bernaards et al., 2025). Consistent with the original conceptualization of the instrument, we primarily evaluated a unidimensional solution, but as a precautionary measure, we also inspected alternative solutions with more than one factor. Decisions about factor retention were guided by a parallel analysis based on factor eigenvalues, in which the eigenvalues obtained from the empirical tetrachoric matrix were compared with those derived from randomly generated data matrices of the same size. Floor/ceiling effects were also evaluated, with values $> 15\%$ indicating possible clustering at the extremes.

In addition, we examined measurement invariance of the general mentalization factor across sex and educational level using multigroup confirmatory factor analysis (CFA) implemented in the lavaan 0.6-21 package (Rosseel, 2012). Building on this CFA model, we re-estimated it in a multigroup framework, treating all items as dichotomous and using probit CFA under the theta parameterization

with the robust Weighted Least Squares with Mean and Variance estimator (WLSMV). The multigroup model syntax was generated using the *semTools* package (Jorgensen et al., 2026), applying the identification strategy for categorical indicators proposed by Wu & Estabrook (2016). For each grouping variable, we then fitted a two-step hierarchy of invariance models. First, a configural model specified the same unidimensional factor structure across groups, while allowing all factor loadings and item thresholds to be freely estimated in each group. Second, a strong invariance model constrained both factor loadings and thresholds to equality across groups for the corresponding grouping variable. Because each item has only one threshold and residual variances are fixed for identification in this probit CFA setting, the joint equality of loadings and thresholds constitutes the practically relevant level of measurement invariance for comparing latent means with binary indicators (Muthén, 2013; Wu & Estabrook, 2016). Model fit was evaluated using chi-square test (χ^2), comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA) with 90% confidence intervals and associated close-fit tests, and standardized root mean square residual (SRMR), following conventional guidelines (Hu & Bentler, 1999). Measurement invariance for sex and educational level was assessed by comparing the configural and strong models using the Satorra-Bentler scaled χ^2 difference test for WLSMV, together with changes in CFI and RMSEA relative to recommended cutoffs for approximate invariance (Chen, 2007; Cheung & Rensvold, 2002).

In accordance with the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), these procedures provide reliability evidence for the MASC-SP scores and validity evidence based on its internal structure and interrelationships. Validity claims refer to the interpretation and use of the MASC-SP scores, rather than to the properties of the test itself (Sireci & Benítez, 2023).

To generate normative data, participants were initially grouped into 5-year age intervals, with exceptions at both ends. A narrower interval was applied at the lower end (18-19 years) to capture possible differences related to the development of SC abilities in young adults, while a wider interval was used at the upper end (50-64 years) to address sample size limitations and improve the stability of normative estimates. To examine whether age-related differences in MASC-SP scores were meaningful for the derivation of age-stratified normative data, Kruskal-Wallis H tests were performed, followed by post hoc comparisons between adjacent groups using Mann-Whitney U tests. To define the final age groups, adjacent age groups that did not show statistically significant differences were progressively combined, resulting in aggregated comparison groups (Figures S1-S4). Mann-Whitney U tests were also performed to examine whether differences in MASC-SP scores by sex and educational level were meaningful for the derivation of normative data stratified by these variables. Where appropriate, Bonferroni corrections were applied. Normative data were stratified according to demographic variables that showed statistically significant differences in Mann-Whitney U tests and are presented as empirical percentiles and z -scores.

Results

Table 1 shows the sample characteristics. Descriptive statistics, Kolmogorov-Smirnov tests, and histograms of MASC-SP scores for all participants and stratified by demographic variables are presented in Tables S1 and S2 and Figures S5-S37.

Table 1
Sample Characteristics

	All participants (N = 2,582)
Female sex	1,804 (69.90%)
Age	29 (18-64)
University education (> 12 years)	1,896 (73.40%)
MASC-SP	
Total score	33 (0-45)
Over-mentalization errors	6 (0-45)
Under-mentalization errors	4 (0-29)
Non-mentalization errors	2 (0-45)

Note. MASC-SP, Movie for the Assessment of Social Cognition-Spanish. Results are presented as n (%) or median (range).

Psychometric Evaluation of MASC-SP Scores

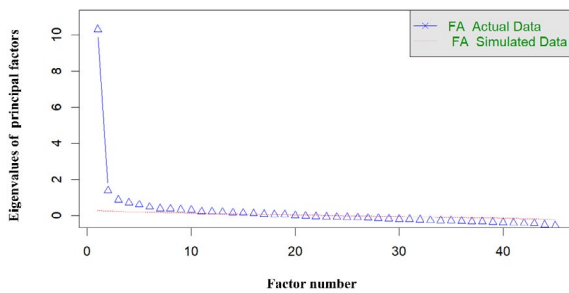
Spearman correlations showed moderate negative associations between the total score and all error scores, a weak positive association between under-mentalization and non-mentalization error scores, a negligible positive association between over-mentalization and non-mentalization error scores, and a non-significant association between over-mentalization and under-mentalization error scores (Table S3).

An EFA was performed for each score. Sampling adequacy was supported by Kaiser-Meyer-Olkin indices ranging from good (.72) to excellent (.91), and data suitability for factor analysis was confirmed by significant Bartlett's tests of sphericity (Table S4). Initial factor extraction suggested multiple factors with eigenvalues > 1 for each score. However, inspection of scree plots consistently indicated a clearly dominant first factor across all scores, characterized by a sharp drop from the first to the second factor, followed by a gradual, nearly linear decline in subsequent eigenvalues (Figures S38-S41). The first factor explained 13.32% of the variance in the total score, 6.83% in the over-mentalization error score, 8.83% in the under-mentalization error score, and 11.79% in the non-mentalization error score. No subsequent factor accounted for > 4.66% of the variance in any score. The item loadings for the first factor of each score are shown in Table S5. Loadings on subsequent factors were generally weak (< .30), often cross-loaded, and with some negative values. Therefore, given the primarily exploratory nature of the study and its applied objectives, retaining a predominantly unifactorial solution for each score was considered the most parsimonious, theoretically consistent, and clearly interpretable approach for normative and clinical purposes.

An additional EFA was performed on the tetrachoric correlation matrix of the total score. The first factor showed an eigenvalue of 10.32, representing 22.90% of the total variance of the items. The remaining factors had substantially smaller eigenvalues (all ≤ 1.38), confirming a marked dominance of the first dimension. To

complement these results, we performed a parallel analysis based on factor eigenvalues, which revealed that although a relatively large number of factors had eigenvalues slightly above those expected under random data, only the first factor clearly exceeded the simulated eigenvalues by a substantial margin. Subsequent factors showed much smaller increments over chance and a gradual, almost linear decline (Figure 2). In line with current recommendations for item-level factor analysis, we therefore retained a predominantly unidimensional solution for the total score, interpreting the first factor as a general mentalizing performance dimension and acknowledging that a considerable proportion of variance remains attributable to item-specific and residual components.

Figure 2
Parallel Analysis of the MASC-SP Total Score (45 Dichotomized Items)



Note. MASC-SP, Movie for the Assessment of Social Cognition-Spanish; FA, Factor analysis.

As a sensitivity check, we also explored a two-factor solution. The first factor accounted for 12.30% of the variance and showed moderate to high loadings for most items. The second factor accounted for an additional 10.60% of the variance, consisting mainly of a small subset of items with substantial cross-loadings and some negative values. This pattern did not yield a theoretically coherent or clearly interpretable multidimensional structure that was also consistent with the original conceptualization of the instrument, and it largely overlapped with the dominant general factor identified in the one-factor solution. Together with the results of the parallel analysis, these findings support retaining a predominantly unidimensional structure for the total score.

Cronbach's alpha values were .80 (total score), .59 (over-mentalization error score), .62 (under-mentalization error score), and .74 (non-mentalization error score).

No floor/ceiling effects were observed in the total score or error scores, except in the case of non-mentalization, which showed a floor effect (20.10%) (Table S6).

Measurement invariance across sex and educational level was evaluated with a two-step sequence of configural and strong models. For sex, the configural model, which specified the same unidimensional factor structure for women and men with freely estimated loadings and thresholds, showed good fit ($\chi^2(1,890) = 3497.80$; CFI = .955; TLI = .953; RMSEA = .026; SRMR = .068), indicating comparable latent structure across groups. Constraining factor loadings and thresholds

to equality (strong invariance) produced some deterioration in fit, although indices still indicated good overall fit ($\chi^2(1,934) = 3973.17$; CFI = .943; TLI = .942; RMSEA = .029; SRMR = .073), and the Satorra-Bentler scaled chi-square difference test was significant, $\Delta\chi^2(44) = 90.71, p < .001$. Even so, changes in approximate fit indices were small ($\Delta\text{CFI} = -.012$; $\Delta\text{RMSEA} = +.003$; $\Delta\text{SRMR} = +.005$) and within recommended cutoffs for approximate invariance, supporting the view that strong invariance across sex is tenable in practical terms and that latent mean comparisons between women and men are substantively interpretable. For educational level, the configural model also showed good fit, suggesting that the unidimensional mentalizing factor operated similarly among university-educated versus non-university-educated participants ($\chi^2(1,890) = 3378.77$; CFI = .953; TLI = .951; RMSEA = .025; SRMR = .066). Imposing equality constraints on factor loadings and thresholds across these groups (strong invariance) again led to some loss of fit, although the global indices remained acceptable ($\chi^2(1,934) = 3481.42$; CFI = .943; TLI = .941; RMSEA = .027; SRMR = .070), and the Satorra-Bentler scaled chi-square difference test indicated significantly worse fit than the configural model, $\Delta\chi^2(44) = 102.65, p < .001$. As in the case of sex, the shifts in approximate fit indices were modest ($\Delta\text{CFI} = -.010$; $\Delta\text{RMSEA} = +.002$; $\Delta\text{SRMR} = +.004$) and stayed within commonly used criteria for approximate invariance, indicating that strong invariance across educational level is also met to a practically acceptable extent and that latent mean differences between education-level groups can be interpreted as reflecting substantive rather than measurement artifacts.

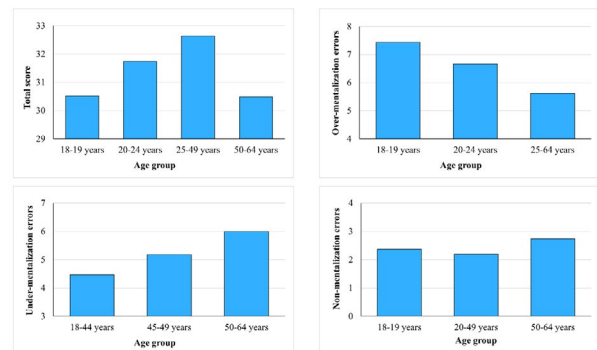
Normative Data for the MASC-SP Total Score

Women scored higher than men (Table S7), and participants with a university education also scored higher than those without (Table S8); this effect was observed in women and men (Table S9).

Total scores differed across initial age groups, $H(7) = 72.91, p < .001$ (Figure S1). In the final age groups, participants aged 18-19 scored lower than those aged 20-24, who in turn scored lower than those aged 25-49. Participants aged 25-49 scored higher than those aged 50-64 (Figure 3, Table S10).

Normative data are presented in Tables 2 and 3.

Figure 3
MASC-SP Scores by Age Group. Means are Reported



Note. MASC-SP, Movie for the Assessment of Social Cognition-Spanish.

Table 2Normative Data for the MASC-SP Total Score in Male Participants by Age Group and Educational Level ($N = 778$)

MASC-SP total score in men											
z-score	PC	Non-university education (≤ 12 years)					University education (> 12 years)				
		18-19	20-24	25-49	50-64	Total	18-19	20-24	25-49	50-64	Total
-2.326	1	10	0	0	-	0	-	1	11	11	11
-1.645	5	13	13	13	15	14	17	22	21	14	20
-1.282	10	19	18	17	18	18	19	27	27	25	27
-.842	20	23	24	-	22	23	27	30	30	28	30
-.524	30	25	28	23	24	25	30	32	31	30	31
-.253	40	27	30	26	25	27	31	-	33	32	33
0	50	29	31	28	27	29	-	33	34	33	-
.253	60	31	33	30	28	31	34	34	35	-	34
.524	70	32	34	33	32	33	35	35	36	34	35
.842	80	33	35	34	34	34	36	36	37	36	37
1.282	90	34	38	36	35	36	37	37	38	38	38
1.645	95	35	39	40	37	38	38	39	39	39	39
2.326	99	-	-	44	-	41	-	-	41	-	41
<i>n</i>		55	90	126	24	295	27	98	308	50	483
<i>M</i>		27.564	29.356	27.587	27.000	28.075	31.148	32.204	32.834	31.420	32.466
<i>SD</i>		6.058	8.102	7.928	6.269	7.560	6.163	5.936	5.532	6.276	5.740
<i>Mdn</i>		29	31	28	26.500	29	34	33	34	33	34
Range		10-37	0-41	0-45	15-38	0-45	17-38	1-42	3-42	11-39	1-42
Skewness		-1.030	-1.480	-.793	-.094	-.986	-1.231	-2.997	-1.806	-1.691	-2.002
Kurtosis		.848	2.755	1.321	-.687	1.596	.601	13.403	5.185	3.131	6.277

Note. MASC-SP, Movie for the Assessment of Social Cognition-Spanish; PC, Percentile; *M*, Mean; *SD*, Standard Deviation; *Mdn*, Median.**Table 3**Normative Data for the MASC-SP Total Score in Female Participants by Age Group and Educational Level ($N = 1,804$)

MASC-SP total score in women											
z-score	PC	Non-university education (≤ 12 years)					University education (> 12 years)				
		18-19	20-24	25-49	50-64	Total	18-19	20-24	25-49	50-64	Total
-2.326	1	13	11	7	9	11	20	13	14	14	14
-1.645	5	18	20	18	13	19	23	25	25	23	24
-1.282	10	22	24	22	16	21	26	27	29	26	27
-.842	20	26	26	27	20	26	28	29	31	29	30
-.524	30	28	29	29	22	28	30	31	32	30	32
-.253	40	30	31	30	28	30	31	32	33	32	33
0	50	31	32	32	29	31	33	33	34	33	34
.253	60	32	33	33	30	33	35	34	35	34	35
.524	70	33	34	34	-	34	36	35	36	35	36
.842	80	36	35	35	33	35	37	36	37	37	37
1.282	90	37	37	37	34	37	-	38	39	38	38
1.645	95	39	38	38	35	38	38	39	40	40	39
2.326	99	-	42	42	-	41	40	41	41	41	41
<i>n</i>		78	105	165	43	391	103	317	890	103	1,413
<i>M</i>		30.026	31.010	30.509	26.884	30.148	32.311	32.511	33.685	32.359	33.225
<i>SD</i>		5.930	5.402	6.152	6.877	6.097	4.627	5.193	4.877	4.974	4.971
<i>Mdn</i>		30.500	32	32	29	31	33	33	34	33	34
Range		13-40	11-42	7-43	9-36	7-43	20-40	9-45	6-45	14-41	6-45
Skewness		-.754	-.753	-1.300	-.777	-1.028	-.552	-1.471	-1.750	-.778	-1.511
Kurtosis		.661	.878	2.267	-.303	1.283	-.434	3.785	5.244	1.049	3.986

Note. MASC-SP, Movie for the Assessment of Social Cognition-Spanish; PC, Percentile; *M*, Mean; *SD*, Standard Deviation; *Mdn*, Median.

Normative Data for the MASC-SP Error Scores

Women made fewer over-mentalization and non-mentalization errors than men, with no significant sex differences in under-mentalization errors (Table S7). Participants with a university education made fewer over-mentalization, under-mentalization, and non-mentalization errors than those without (Table S8). The effect of educational level on over-mentalization and non-mentalization errors was observed in women and men (Table S9).

Over-mentalization errors differed across initial age groups, $H(7) = 106.50, p < .001$ (Figure S2). In the final age groups, participants aged 18-19 made more over-mentalization errors than those aged 20-24, who in turn made more than those aged 25-64 (Figure 3, Table S10).

Under-mentalization errors differed across initial age groups, $H(7) = 70.67, p < .001$ (Figure S3). In the final age groups, participants aged 18-44 made fewer under-mentalization errors than those aged 45-49, who in turn made fewer than those aged 50-64 (Figure 3, Table S10).

Non-mentalization errors differed across initial age groups, $H(7) = 28.50, p < .001$ (Figure S4). In the final age groups, participants aged 18-19 made more non-mentalization errors than those aged 20-49, who in turn made fewer than those aged 50-64 (Figure 3, Table S10).

Normative data are presented in Tables S11-S15.

Discussion

This study presents the first large-scale psychometric evaluation and normative framework for MASC-SP scores in Spanish-speaking adults. For the first time, the psychometric properties of total and error scores have been evaluated. The results support a predominantly unifactorial structure for each score, with good internal consistency for the total score, acceptable internal consistency for the non-mentalization error score, and lower but still interpretable internal consistency for the over-mentalization and under-mentalization error scores. Based on this psychometric foundation, normative data are also presented stratified by sex and educational level, which were shown to function under conditions of measurement invariance.

Spearman correlations showed negative associations between the total score and all error scores, along with positive and selective correlations between error scores, thus capturing partially overlapping but not redundant aspects of general mentalization performance and associated error patterns, in line with the original conceptualization of the instrument's scores (Dziobek et al., 2006; Lahera et al., 2014).

Although the proportion of variance explained by the one-factor solution was modest, subsequent factors accounted for much smaller amounts of variance and did not form coherent alternative structures. In this regard, the tetrachoric correlation matrix of the total score, together with the parallel analysis, provided converging evidence that only the first factor clearly exceeded random eigenvalues, supporting the retention of a predominantly unidimensional solution.

Consistent with this structure, the total score showed the highest reliability, supporting its use as a robust primary quantitative indicator of general mentalization performance in Spanish-speaking adults. In contrast, error scores showed more heterogeneous reliability, supporting their interpretation as complementary indicators of specific response patterns, rather than as scores reflecting independent latent

constructs, in line with their original conceptualization as secondary outcomes (Dziobek et al., 2006; Lahera et al., 2014).

The ceiling effect commonly observed in scores on traditional SC instruments (Pinkham et al., 2018) was not observed in our results, but a floor effect was detected in the non-mentalization error score, as one-fifth of participants, particularly those with university education, made few or no non-mentalization errors. This is consistent with results obtained in non-clinical samples characterized by intact mentalization abilities (Cotter et al., 2018) and supports that MASC-SP scores provide an adequate range of measurements and are sensitive to individual differences in healthy populations, especially beyond the lower end of the error distribution and within the mid-to-high range of error frequency. Importantly, greater discriminatory power of the non-mentalization error score is expected in clinical samples, whose error rates are typically higher than those of healthy populations (Dziobek et al., 2006; Lahera et al., 2014).

Overall, these results are consistent with those of previous psychometric studies, which have reported a comparable unifactorial solution (Chang et al., 2024; Fossati et al., 2018; Müller et al., 2016) and similar internal consistency for the total score (Chang et al., 2024; Dziobek et al., 2006; Fossati et al., 2018; Huang et al., 2023; Lahera et al., 2014; Martinez et al., 2017; Sarig-Shmueli et al., 2023). However, none of these studies examined the factor structure, reliability, and distributional properties of error scores, highlighting the novelty of the present work. These analyses were exploratory and did not aim to establish formal latent trait models, but rather to evaluate whether each error response category was characterized by a predominant source of shared variance, so that each error score could be considered a distinguishable and interpretable aspect of instrument performance beyond the total score.

This study also presents the first normative framework for MASC-SP scores using empirical percentiles and *z*-scores. The data were derived from a large-scale community sample and are applicable to research and clinical settings. Sex differences showed small-magnitude effect sizes, with women scoring higher overall and making fewer over-mentalization and non-mentalization errors than men. Educational level showed small-to-medium effect sizes, with university-educated participants scoring higher overall and making fewer mentalization errors of any kind than non-university-educated participants. Importantly, measurement invariance across sex and educational level was confirmed, indicating that differences between groups likely reflect substantive variations in general mentalization performance rather than measurement artifacts, justifying the development of group-stratified norms.

In contrast, the influence of age was heterogeneous. While total scores varied across age groups, with middle-aged participants scoring higher than younger and older groups, non-mentalization errors showed an inverse pattern, mirroring differences in total scores. Furthermore, over-mentalization errors tended to decrease with age, and under-mentalization errors tended to increase. Despite these results, differences between adjacent age groups were characterized by small-to-negligible effect sizes, calling into question their relevance as clinically meaningful or developmentally related changes. Consequently, age-stratified norms are presented for descriptive purposes only and should be used with caution. Therefore, it is recommended that non-age-stratified scores be prioritized over age-stratified scores when applying these norms in

research and clinical settings. Further studies are needed to provide continuous norms using age-conditioned metrics to reduce potential artifacts introduced by age binning.

Overall, these results are consistent with those of the Spanish adaptation of the MASC, in which healthy participants showed a similar scoring profile to that of the participants in the present study (Lahera et al., 2014), and with those of other studies that have examined the influence of sex and educational level on SC performance using traditional instruments (Christov-Moore et al., 2014; Navarra-Ventura et al., 2018, 2021, 2022; Vaskinn et al., 2024). The results are also consistent with recent evidence that challenges the traditional view of age-related cognitive change as a linear process and instead emphasizes that mentalizing performance across the lifespan is determined by a heterogeneous set of factors, including individual differences in cognitive/executive functions, brain reserve, and brain pathology (Arioli et al., 2018; Fittipaldi et al., 2024). In summary, when applying these norms, it is essential to bear in mind that educational level had a consistent influence on general mentalizing performance and associated error patterns, while sex only influenced specific scores. Age, on the other hand, exerted influences whose clinical or developmental significance is questionable.

When evaluating the strengths and weaknesses of normative data, several criteria have been proposed (Mitrushina et al., 2005): (1) ≥ 50 participants per cell; detailed reports on (2) sample composition and (3) test administration; presentation of data by (4) sex, (5) age, and (6) education/intelligence quotient (IQ); and (7) presentation of descriptive statistics for each normative cell. The present study meets five of these criteria. The absence of IQ data (criterion 6) is unlikely to be a major limitation, given the stratification of data by educational level and the well-established positive correlation between education and IQ (Deary & Johnson, 2010). A second limitation is that not all normative cells included ≥ 50 participants (criterion 1), particularly those corresponding to non-university-educated male participants in the younger and older age groups. This shortcoming is mitigated by the total sample size, which is larger than that of previous studies on the MASC and MASC-SP (Chang et al., 2024; Dziobek et al., 2006; Fossati et al., 2018; Huang et al., 2023; Lahera et al., 2014; Martínez et al., 2017; Müller et al., 2016; Sarig-Shmueli et al., 2023). Furthermore, when normative cells are considered independently of age groups, all of them meet the threshold of ≥ 50 participants, reinforcing the recommendation to prioritize non-age-stratified scores over age-stratified scores when using the norms. A third limitation concerns the use of non-probability sampling and voluntary online participation, which may have generated a certain degree of self-selection bias, resulting in a sample with a sociodemographic profile skewed toward university-educated young women. This limits the generalizability of the results and suggests that normative data should be considered preliminary and more representative of populations with sociodemographic characteristics similar to those in the present study. Finally, no information was collected on the region of residence of participants within Spain, which prevented reliable alignment of the norms with population benchmarks in key geographic strata. Future research using probabilistic and stratified sampling methods is needed to refine these norms and establish region-specific validations of the instrument scores. This is of great importance, especially for Latin American countries, as cultural, contextual, and linguistic factors may influence mentalization performance (Fittipaldi et al., 2024), so the direct application of these Spain-based norms to other Spanish-speaking populations outside Spain should be approached with caution.

In conclusion, this study presents the first large-scale psychometric evaluation and normative framework for MASC-SP scores in Spanish-speaking adults. The results provide reliability evidence for total and error scores, as well as validity evidence based on internal structure and interrelationships, supporting the interpretation of these scores as indicators of general mentalization performance and associated error patterns. Within this framework, the total score can be used as a robust primary quantitative indicator of general mentalization performance, while error scores are best understood as secondary qualitative indicators of specific response patterns. Normative data stratified by sex and educational level provide a useful reference for describing an individual's position relative to demographically comparable reference groups and for making group-level comparisons in research and clinical settings. However, the validity evidence presented is mainly based on internal structure and does not yet support higher-level uses such as individual diagnostic classification or treatment allocation. Extending the interpretation of these scores to such uses requires accumulating additional sources of validity evidence, particularly evidence based on relationships with other variables (Dziobek et al., 2006; Lahera et al., 2014), on response processes, and on testing consequences (American Educational Research Association et al., 2014; Sireci & Benítez, 2023).

Author Contributions

Aina Llompарт-Casasnovas: Data curation, Formal analysis, Visualization, Writing-Review & Editing. **Adoración Castro:** Writing-Review & Editing. **Margalida Gili:** Supervision, Writing-Review & Editing. **Guillem Navarra-Ventura:** Conceptualization, Data curation, Formal analysis, Project administration, Visualization, Writing-Original Draft, Writing-Review & Editing. **Roberto Rodríguez-Jimenez:** Funding acquisition, Investigation, Writing-Review & Editing. **Miguel A. Ortega:** Investigation, Writing-Review & Editing. **Leticia Boada:** Investigation, Writing-Review & Editing. **Mercè Jodar:** Writing-Review & Editing. **Narcís Cardoner:** Writing-Review & Editing. **Albert Sesé:** Formal analysis, Methodology, Writing-Review & Editing. **Miquel Roca:** Supervision, Writing-Review & Editing. **Guillermo Lahera:** Conceptualization, Funding acquisition, Investigation, Resources, Writing-Review & Editing.

Funding

This work was funded by the Instituto de Salud Carlos III (ISCIII) and co-funded by the European Regional Development Fund (ERDF) (grant number PI21/01252). Additional support was provided by the Regional Government of Madrid (R&D Program in Biomedicine, S2022/BMD-7216, AGES 3-CM) and by the Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM, ISCIII).

The source of funding did not participate in the design of the study, the data collection, analysis, or interpretation, the writing of the article, or in the decision to submit it for publication.

Declaration of Interest

Guillem Navarra-Ventura, Roberto Rodríguez-Jimenez, Miquel Roca and Guillermo Lahera have served as consultants for,

participated in sponsored activities of, or received research grants from JanssenCilag, Lundbeck, Otsuka, Pfizer, Ferrer, Juste, Takeda, Esteve, Exeltis, Casen Recordati, Angelini, Adamed, Boehringer, or Rovi. The remaining authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request. Supplementary materials are accessible at https://osf.io/46gij/?view_only=ea167e207bcf4c319f41e2b19bdbcb23

References

- Aidelbaum, R., & Goghari, V. (2022). A visual task-based assessment of theory of mind and social perception within bipolar disorder. *Journal of Nervous & Mental Disease*, 210(1), 37–44. <https://doi.org/10.1097/NMD.0000000000001408>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Andrade-González, N., Sarasa, M., García-López, A., Leonés, I., Halverson, T. F., & Lahera, G. (2021). Mentalizing errors in patients with schizophrenia who received psychosocial rehabilitation: A case-control study. *Psychiatric Quarterly*, 92(3), 947–959. <https://doi.org/10.1007/s11126-020-09863-x>
- Andreou, C., Kelm, L., Bierbrodt, J., Braun, V., Lipp, M., Yassari, A. H., & Moritz, S. (2015). Factors contributing to social cognition impairment in borderline personality disorder and schizophrenia. *Psychiatry Research*, 229(3), 872–879. <https://doi.org/10.1016/j.psychres.2015.07.057>
- Arioli, M., Crespi, C., & Canessa, N. (2018). Social cognition through the lens of cognitive and clinical neuroscience. *BioMed Research International*, 2018, Article 4283427. <https://doi.org/10.1155/2018/4283427>
- 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.
- Bernaards, C. A., Gilbert, P., & Jennrich, R. I. (2025). *GPArotation: Gradient projection factor rotation (Version 2025.3-1)* [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=GPArotation>
- Catalan, A., Angosto, V., Díaz, A., Martínez, N., Guede, D., Pereda, M., Madrazo, A., Bustamante, S., Bilbao, A., Osa, L., Inchausti, L., & Gonzalez-Torres, M. A. (2018). The relationship between theory of mind deficits and neurocognition in first episode-psychosis. *Psychiatry Research*, 268, 361–367. <https://doi.org/10.1016/j.psychres.2018.06.066>
- Cavieres, A., Acuña, V., Arancibia, M., & Escobar, C. (2025). Advances in the ecological validity of research on social cognition in schizophrenia: A systematic review of the literature. *Schizophrenia Research: Cognition*, 39, Article 100333. <https://doi.org/10.1016/j.scog.2024.100333>
- Chang, Y., Yu, C., Huang, C., Wang, T., Dziobek, I., & Lane, H. (2024). Discrepancy of social cognition between bipolar disorders and major depressive disorders. *Brain and Behavior*, 14(1), Article e3365. <https://doi.org/10.1002/brb3.3365>
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5
- Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacoboni, M., & Ferrari, P. F. (2014). Empathy: Gender effects in brain and behavior. *Neuroscience & Biobehavioral Reviews*, 46, 604–627. <https://doi.org/10.1016/j.neubiorev.2014.09.001>
- Cotter, J., Granger, K., Backx, R., Hobbs, M., Looi, C. Y., & Barnett, J. H. (2018). Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neuroscience & Biobehavioral Reviews*, 84, 92–99. <https://doi.org/10.1016/j.neubiorev.2017.11.014>
- Deary, I. J., & Johnson, W. (2010). Intelligence and education: Causal perceptions drive analytic processes and therefore conclusions. *International Journal of Epidemiology*, 39(5), 1362–1369. <https://doi.org/10.1093/ije/dyq072>
- Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., Kessler, J., Woike, J. K., Wolf, O. T., & Convit, A. (2006). Introducing MASC: A Movie for the Assessment of Social Cognition. *Journal of Autism and Developmental Disorders*, 36(5), 623–636. <https://doi.org/10.1007/s10803-006-0107-0>
- Ekman, P., & Friesen, W. C. (1976). *Pictures of facial affect*. Consulting Psychologists Press.
- Engelstad, K. N., Rund, B. R., Torgalsbøen, A.-K., Lau, B., Ueland, T., & Vaskinn, A. (2019). Large social cognitive impairments characterize homicide offenders with schizophrenia. *Psychiatry Research*, 272, 209–215. <https://doi.org/10.1016/j.psychres.2018.12.087>
- Fittipaldi, S., Legaz, A., Maito, M., Hernandez, H., Altschuler, F., Canziani, V., Moguilner, S., Gillan, C. M., Castillo, J., Lillo, P., Custodio, N., Avila-Funes, J. A., Cardona, J. F., Slachevsky, A., Henriquez, F., Fraile-Vazquez, M., Cruz de Souza, L., Borroni, B., Hornberger, M., ... Ibanez, A. (2024). Heterogeneous factors influence social cognition across diverse settings in brain health and age-related diseases. *Nature Mental Health*, 2(1), 63–75. <https://doi.org/10.1038/s44220-023-00164-3>
- Fossati, A., Borroni, S., Dziobek, I., Fonagy, P., & Somma, A. (2018). Thinking about assessment: Further evidence of the validity of the Movie for the Assessment of Social Cognition as a measure of mentalistic abilities. *Psychoanalytic Psychology*, 35(1), 127–141. <https://doi.org/10.1037/pap0000130>
- Fretland, R. A., Andersson, S., Sundet, K., Andreassen, O. A., Melle, I., & Vaskinn, A. (2015). Theory of mind in schizophrenia: Error types and associations with symptoms. *Schizophrenia Research*, 162(1–3), 42–46. <https://doi.org/10.1016/j.schres.2015.01.024>
- Gur, R. C. (2002). Sex differences in temporo-limbic and frontal brain volumes of healthy adults. *Cerebral Cortex*, 12(9), 998–1003. <https://doi.org/10.1093/cercor/12.9.998>
- Happé, F. G. (1994). An advanced test of theory of mind: Understanding of story characters’ thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, 24(2), 129–154. <https://doi.org/10.1007/BF02172093>

- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. <https://doi.org/10.1080/10705519909540118>
- Huang, Y.-L., Chen, T.-T., Dziobek, I., & Tseng, H.-H. (2023). Mentalizing in a Movie for the Assessment of Social Cognition (MASC): The validation in a taiwanese sample. *BMC Psychology*, 11(1), Article 287. <https://doi.org/10.1186/s40359-023-01321-0>
- Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Hakonarson, H., Gur, R. E., Gur, R. C., & Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2), 823-828. <https://doi.org/10.1073/pnas.1316909110>
- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2026). *semTools: Useful tools for structural equation modeling (Versión 0.5-8)*. [R package]. R Foundation for Statistical Computing. <https://CRAN.R-project.org/package=semTools>
- Lahera, G., Boada, L., Pousa, E., Mirapeix, I., Morón-Nozaleda, G., Marinas, L., Gisbert, L., Pamiàs, M., & Parellada, M. (2014). Movie for the Assessment of Social Cognition (MASC): Spanish validation. *Journal of Autism and Developmental Disorders*, 44(8), 1886-1896. <https://doi.org/10.1007/s10803-014-2061-6>
- Martinez, G., Alexandre, C., Mam-Lam-Fook, C., Bendjema, N., Gaillard, R., Garel, P., Dziobek, I., Amado, I., & Krebs, M. O. (2017). Phenotypic continuum between autism and schizophrenia: Evidence from the Movie for the Assessment of Social Cognition (MASC). *Schizophrenia Research*, 185, 161-166. <https://doi.org/10.1016/j.schres.2017.01.012>
- Mitchell, R. L. C., & Phillips, L. H. (2015). The overlapping relationship between emotion perception and theory of mind. *Neuropsychologia*, 70, 1-10. <https://doi.org/10.1016/j.neuropsychologia.2015.02.018>
- Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment*. Oxford University Press.
- Montag, C., Dziobek, I., Richter, I. S., Neuhaus, K., Lehmann, A., Sylla, R., Heekeren, H. R., Heinz, A., & Gallinat, J. (2011). Different aspects of theory of mind in paranoid schizophrenia: Evidence from a video-based assessment. *Psychiatry Research*, 186(2-3), 203-209. <https://doi.org/10.1016/j.psychres.2010.09.006>
- Montag, C., Ehrlich, A., Neuhaus, K., Dziobek, I., Heekeren, H. R., Heinz, A., & Gallinat, J. (2010). Theory of mind impairments in euthymic bipolar patients. *Journal of Affective Disorders*, 123(1-3), 264-269. <https://doi.org/10.1016/j.jad.2009.08.017>
- Msika, E.-F., Despres, M., Piolino, P., & Narme, P. (2024). Dynamic and/or multimodal assessments for social cognition in neuropsychology: Results from a systematic literature review. *The Clinical Neuropsychologist*, 38(4), 922-962. <https://doi.org/10.1080/13854046.2023.2266172>
- Müller, N., Baumeister, S., Dziobek, I., Banaschewski, T., & Poustka, L. (2016). Validation of the Movie for the Assessment of Social Cognition in adolescents with ASD: Fixation duration and pupil dilation as predictors of performance. *Journal of Autism and Developmental Disorders*, 46(9), 2831-2844. <https://doi.org/10.1007/s10803-016-2828-z>
- Muthén, B. (2013). *Binary item measurement invariance testing with categorical factor analysis* [Unpublished technical report]. Retrieved from <https://www.statmodel.com/download/Binary%20item%20measurement%20invariance%20testing.pdf>
- Navarra-Ventura, G., Fernandez-Gonzalo, S., Turon, M., Pousa, E., Palao, D., Cardoner, N., & Jodar, M. (2018). Gender differences in social cognition: A cross-sectional pilot study of recently diagnosed patients with schizophrenia and healthy subjects. *Canadian Journal of Psychiatry*, 63(8), 538-546. <https://doi.org/10.1177/0706743717746661>
- Navarra-Ventura, G., Vicent-Gil, M., Serra-Blasco, M., Cobo, J., Fernández-Gonzalo, S., Goldberg, X., Jodar, M., Crosas, J. M., Palao, D., Lahera, G., Vieta, E., & Cardoner, N. (2022). Higher order theory of mind in patients with bipolar disorder and schizophrenia/schizoaffective disorder. *European Archives of Psychiatry and Clinical Neuroscience*, 272(3), 497-507. <https://doi.org/10.1007/s00406-021-01265-9>
- Navarra-Ventura, G., Vicent-Gil, M., Serra-Blasco, M., Massons, C., Crosas, J. M., Cobo, J., Jubert, A., Jodar, M., Fernández-Gonzalo, S., Goldberg, X., Palao, D., Lahera, G., Vieta, E., & Cardoner, N. (2021). Group and sex differences in social cognition in bipolar disorder, schizophrenia/schizoaffective disorder and healthy people. *Comprehensive Psychiatry*, 109, Article 152258. <https://doi.org/10.1016/j.comppsy.2021.152258>
- Newbury-Helps, J., Feigenbaum, J., & Fonagy, P. (2017). Offenders with antisocial personality disorder display more impairments in mentalizing. *Journal of Personality Disorders*, 31(2), 232-255. <https://doi.org/10.1521/pedi.2016.30.246>
- Pinkham, A. E., Harvey, P. D., & Penn, D. L. (2018). Social cognition psychometric evaluation: Results of the final validation study. *Schizophrenia Bulletin*, 44(4), 737-748. <https://doi.org/10.1093/schbul/sbx117>
- Preißler, S., Dziobek, I., Ritter, K., Heekeren, H. R., & Roepke, S. (2010). Social cognition in borderline personality disorder: Evidence for disturbed recognition of the emotions, thoughts, and intentions of others. *Frontiers in Behavioral Neuroscience*, 4, Article 182. <https://doi.org/10.3389/fnbeh.2010.00182>
- Quek, J., Bennett, C., Melvin, G. A., Saeedi, N., Gordon, M. S., & Newman, L. K. (2018). An investigation of the mentalization-based model of borderline pathology in adolescents. *Comprehensive Psychiatry*, 84, 87-94. <https://doi.org/10.1016/j.comppsy.2018.04.005>
- Revelle, W. (2025). *psych: Procedures for psychological, psychometric, and personality research (Version 2.5.3)* [R package]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Romero-Ferreiro, V., García-Fernández, L., Aparicio, A. I., Martínez-Gras, I., Dompablo, M., Sánchez-Pastor, L., Rentero, D., Alvarez-Mon, M. Á., Espejo-Saavedra, J. M., Lahera, G., Mari-Beffa, P., Santos, J. L., & Rodríguez-Jimenez, R. (2022). Emotional processing profile in patients with first episode schizophrenia: The influence of neurocognition. *Journal of Clinical Medicine*, 11(7), Article 2044. <https://doi.org/10.3390/jcm11072044>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1-36. <https://doi.org/10.18637/jss.v048.i02>
- Sahl, A. C., Rognlien, H. F., Andreassen, O. A., Melle, I., Ueland, T., & Vaskinn, A. (2023). Theory of mind in schizophrenia: A comparison of subgroups with low and high IQ. *Nordic Journal of Psychiatry*, 77(4), 329-335. <https://doi.org/10.1080/08039488.2022.2106512>
- Santos, J. M., Pousa, E., Soto, E., Comes, A., Roura, P., Arrufat, F. X., & Obiols, J. E. (2017). Theory of mind in euthymic bipolar patients and first-degree relatives. *The Journal of Nervous and Mental Disease*, 205(3), 207-212. <https://doi.org/10.1097/NMD.0000000000000595>

- Sarig-Shmueli, Y., Kurman, J., Talmon, S., Yoeli, N., Gur, E., Stein, D., & Rothschild-Yakar, L. (2023). Mentalizing ability, mentalizing impairments, and anorexia nervosa: Validation of the hebrew version of the Movie for the Assessment of Social Cognition (MASC). *Clinical Psychology & Psychotherapy*, 30(4), 873–881. <https://doi.org/10.1002/cpp.2848>
- Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., & Kanske, P. (2021). Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, 147(3), 293–327. <https://doi.org/10.1037/bul0000303>
- Seitz, K. I., Ehler, N., Schmitz, M., Schmitz, S. E., Dziobek, I., Herpertz, S. C., & Bertsch, K. (2022). Affective and cognitive theory of mind in posttraumatic stress, major depressive, and somatic symptom disorders: Association with childhood trauma. *British Journal of Clinical Psychology*, 61(3), 680–700. <https://doi.org/10.1111/bjc.12357>
- Sireci, S. G., & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema*, 35(3), 217–226. <https://doi.org/10.7334/psicothema2022.477>
- Somma, A., Ferrara, M., Terronni, A., Frau, C., Ardizzone, I., Sharp, C., & Fossati, A. (2019). Hypermentalizing as a marker of borderline personality disorder in italian adolescents: A cross-cultural replication of Sharp and colleagues' (2011) findings. *Borderline Personality Disorder and Emotion Dysregulation*, 6(1), Article 5. <https://doi.org/10.1186/s40479-019-0104-5>
- Vaskinn, A., Andersson, S., Østefjells, T., Andreassen, O. A., & Sundet, K. (2018). Emotion perception, non-social cognition and symptoms as predictors of theory of mind in schizophrenia. *Comprehensive Psychiatry*, 85, 1–7. <https://doi.org/10.1016/j.comppsy.2018.05.002>
- Vaskinn, A., Ueland, T., Melle, I., & Sundet, K. (2024). Sex differences in social cognition among individuals with schizophrenia and in healthy control participants: A secondary analysis of published data. *Archives of Women's Mental Health*, 27(5), 661–667. <https://doi.org/10.1007/s00737-024-01422-8>
- Wolkenstein, L., Schönenberg, M., Schirm, E., & Hautzinger, M. (2011). I can see what you feel, but I can't deal with it: Impaired theory of mind in depression. *Journal of Affective Disorders*, 132(1-2), 104–111. <https://doi.org/10.1016/j.jad.2011.02.010>
- Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, 81(4), 1014–1045. <https://doi.org/10.1007/s11336-016-9506-0>
- Zwick, J. C., & Wolkenstein, L. (2017). Facial emotion recognition, theory of mind and the role of facial mimicry in depression. *Journal of Affective Disorders*, 210, 90–99. <https://doi.org/10.1016/j.jad.2016.12.022>