

Online Screening for Social Cognition

Rebecca Johannessen

rebecca.johannessen@klinikklangg.ch

Swiss Epilepsy Center

Marcel Eicher

Swiss Epilepsy Center

Hennric Jokeit

Swiss Epilepsy Center

Article

Keywords:

Posted Date: September 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10293801/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Impairments in social cognition are prevalent in neurological and psychiatric conditions, impacting social functioning and serving as potential transdiagnostic markers. Despite being recognised for over a decade in the DSM-5, clinical implementation of assessment is limited due to a lack of well-validated, transculturally applicable, and ecologically valid tests. We developed COSIMO, a 5-minute online tool for emotion recognition, using demographically diverse and ecologically valid video stimuli. We evaluated COSIMO in a cross-sectional validation and normative study. In a German-speaking sample ($n = 2299$) we found diagnostic accuracy with 52–61% sensitivity and 67–84% specificity. Classifier performance was robust for autism spectrum disorder ($AUC = 74\%$) and temporal lobe epilepsy ($AUC = 68\%$) and expectedly lower for ADHD ($AUC = 58\%$). COSIMO showed a low correlation with general cognitive impairment ($\tau = 0.27$, $p = .064$), confirming discriminant validity, and medium correlations with existing social cognitive tests of related constructs ($\tau = 0.34$ – 0.50), supporting convergent validity. Reliability ($r = 0.45$ – 0.56) reflected the trade-off between test brevity and complexity, without compromising clinical validity. COSIMO offers a time-efficient solution for social cognition screening and is available in multiple languages for research and clinical use.

Introduction

Impairments in social cognitive abilities are prevalent among both neurological and psychiatric patient groups^{1,2}, often resulting in diminished subjective quality of life, weakened social support networks, and a heightened risk of comorbidities³. Recognised as a distinct cognitive domain in the DSM-5⁴, social cognition remains significantly under-represented in routine neuropsychological assessments^{5,6}, partly due to the scarcity of informative, ecologically valid, and cost-effective tests^{1,5,7,8}. Recently, literature on social cognition in clinical populations has highlighted the need for more ecologically valid methods to detect social cognitive impairment and a shift towards digital diagnostic tools^{1,5,8}. This study examines the psychometric properties of the digital screening tool, Cognition of Social Interaction in Movies (COSIMO)², which, to our knowledge, is the first online screening tool for complex emotion recognition (CER) designed to address this gap.

Ecological validity is crucial for examining social cognitive impairments that affect daily functioning⁷. For example, higher ecological validity, such as verisimilitude (i.e. similarity of test characteristics to everyday demands in social interactions), arguably enhances a test's ability to reveal clinically relevant dysfunctions in the everyday social lives of patients⁵. However, many social cognition tests with high ecological validity demand a high cognitive load, rendering the results difficult to interpret in neurological patients with impairments in multiple cognitive domains. Consequently, and because many ecologically valid tests are too time-consuming for clinical practice, simpler and artificial social cognitive tests using static images or text-based vignettes are often preferred^{1,5,6,10,11}. While these tests are user-friendly and demonstrate high sensitivity in certain populations, they have been criticised for lacking ecological validity and cultural fairness^{12–15}.

In neuropsychological performance testing, two established social cognition constructs are emotion recognition (ER) and theory of mind (ToM). ER tasks are based on concrete socially salient sensory input and often involve static or brief dynamic images of cropped faces or prosodic voices, requiring participants to identify the correct basic emotion^{12–15}. In contrast, ToM is assessed through tasks with more complex socially salient sensory input and abstract information, requiring higher-order social reasoning or social perception processes to deduce the mental states of others. These domains intersect at the level of perception and deduction of the emotional states of others¹⁶. Complex emotion recognition (CER) serves as a bridge between ToM, which often requires abstract and perceptually detached reasoning, and ER, which is grounded in the processing of socially salient sensory input. However, it extends beyond basic ER by increasing the complexity of both the perceptual information to be integrated and the emotional states to be inferred² (Fig. 1). The term CER was first introduced in a study by Golan et al., which aimed to validate the video-based social cognition test Reading the Mind in Films¹⁶, and partially overlaps with the concept of affective ToM^{7–9,20}. A recent review described complex emotions as those that are longer-lasting and evolving over time, while having significant cognitive appraisal requirements and being dependent on culture⁸. Dynamic CER tasks offer the potential for high ecological validity (i.e. verisimilitude) and consequently greater sensitivity to impairment²⁰.

Advances in digital technology have expanded the opportunities for neuropsychological test development⁸. Screen-based paradigms allow the use of standardised, sensory-rich, and naturalistic dynamic stimuli, making them well-suited for capturing the complexity of social cognition with greater ecological validity⁶. While digital tools are increasingly used for stimulus presentation, their potential for

large-scale online data collection in item development and norming remains underutilised^{21–23}. Online data collection allows access to larger and more diverse samples across cultures while also offering the potential to address cultural and linguistic specificities that shape social cognition, improving its usability for cross-cultural and transdiagnostic purposes²⁴. Additionally, online administration in a browser and automatic scoring make the test easily accessible.

Recent meta-analyses suggest that social cognition is impaired in a broad range of psychiatric and neurological disorders. On the neurodevelopmental side, autism spectrum disorder (ASD) has received much attention^{25–27}, with deficits in social communication and interaction as part of the defining symptoms²⁶. In neurology, temporal lobe epilepsy (TLE) has received increasing attention over the past two decades^{28–31}, as imaging and neurophysiological studies have asserted the importance of both temporal^{32–34} and mesial³⁵ brain structures in the successful interpretation of emotions and social signals. Frontal brain regions are involved as well^{23,36}, and thus, reduced social cognitive abilities have also been found in groups with attention deficit hyperactivity disorder (ADHD)^{25,37,38}, often scoring somewhere between healthy controls and ASD in ER and ToM paradigms^{1,39}. This relationship was also found in a study exploring the effects of ASD, TLE, ADHD, multiple sclerosis (MS), and healthy aging on CER using COSIMO². The recognition of widespread cortical regions underlying social cognition has shifted the field away from localisation theories towards network theories of social cognitive functions⁴¹, which better explain the widely distributed impairments found across the transdiagnostic landscape⁴².

COSIMO is a brief online screening tool with a low cognitive load, high ecological validity, and high item complexity. We developed COSIMO to address the shortcomings identified in existing social cognition assessment methods (listed in Table 1) to increase its feasibility for clinical use. The test was developed by collecting a set of 600 muted, brief video clips of dyadic interactions from existing film material (following the concept of fair use⁴³) and allowing online raters to choose the most appropriate emotion to describe the character in each video clip. In subsequent runs, the videos and options were reduced until we were left with 50 videos with three response options each, which were made available for clinical use and research as two pilot versions with 25 items each. We collected norm- and patient data in mixed online and onsite settings to explore the psychometric properties of COSIMO and to select the best version. This study discusses the validation, norming, and psychometric properties of the final version.

Table 1

Comparison of validity and feasibility factors in the social cognition tests we compared with COSIMO and how we addressed these issues with COSIMO.

Ecological validity	MASC	RMET	FPT	COSIMO
	✓	X	X	✓ - Dynamic videos with high complexity
Economic viability	X	✓	(✓)	✓ - Free to use - Time efficient (~ 5 minutes) - Automatic scoring - Runs on any device in the browser
Cultural fairness	X	X	X	✓ - Ethnically diverse stimuli - Easily translatable
Norms	X	✓	X	✓ - Continuous norming approach - Norm models directly implemented in the app
Cognitive load	X	(✓)	(✓)	✓ - Short stimuli - Simple format
Language requirements	X	(✓)	X	✓ - One-word responses - Only most common synonyms used
Availability	X	✓	✓	✓ - Free to use - Access in any browser
Note. ✓ = satisfied, (✓) = partially satisfied, X = not satisfied, MASC = Movie for the Assessment of Social Cognition, RMET = Reading the Mind in the Eyes Test, FPT = Faux Pas Test, COSIMO = Cognition of Social Interaction in Movies.				

In the current study, we investigated the norm group (healthy control participants) and clinical groups of ASD, TLE, and ADHD, as well as a collective group of all other patient cases aggregated from the app, regardless of diagnosis in a rest category ("other diagnosis"). To establish normative reference values and evaluate the psychometric properties of COSIMO in more detail, we aimed to: a) generate and validate standardised, adjusted norm models for the German version; b) investigate the sensitivity of COSIMO to impaired performance in social cognition as measured by other neuropsychological tests, ASD diagnosis, TLE diagnosis, or ADHD diagnosis using receiver operating curve (ROC) plots and density plots; c) conceptually position COSIMO within the social cognition nomenclature by comparing it with other well-established tests; d) compare the English and German norms to investigate potential language effects; and e) examine reliability. For these purposes, we extracted all existing data from English- and German-speaking participants to examine their properties.

Results

Sample characteristics

The sample consisted of N = 2299 participants, of whom 700 were included in the norm group, 36 were diagnosed with TLE, 115 had a reported ASD diagnosis, and 71 had ADHD. The remaining aggregated results from 1377 patients with other or unknown diagnostic characteristics were included as a rest category (“other”) to compare whether neurological or psychiatric diagnoses, in general, predicted a lower COSIMO score. 37 participants with ASD also reported ADHD but were only included in the ASD group as this was considered the primary diagnosis. An overview of the known diagnoses in the “other” category, as well as all known comorbidities, is provided in the Supplementary Material. Participants either took part online (n = 789) or at a clinic (n = 1510), either as volunteers or during routine neuropsychological assessment. A full description of the distribution across demographic and diagnostic groups is presented in **Table 2**.

Descriptive statistics

In COSIMO, the norm group achieved the highest mean score ($m = 16.23 \pm 1.93$), followed by ADHD ($m = 15.82 \pm 2.24$), other diagnoses ($m = 15.33 \pm 2.41$), TLE ($m = 14.81 \pm 1.85$), and ASD ($m = 14.28 \pm 2.81$). All test results are summarised in **Table 3**, along with the available sample sizes for each test. We investigated the differences in scores between different age, gender, and education groups using two-sided Mann-Whitney-Wilcoxon tests with Holm-Bonferroni corrections. Although the gender effect was not significant, there was a trend towards a lower median score in the male group (male $m = 16.20 \pm 2.04$, $mdn = 16$, female $m = 16.48 \pm 1.88$, $mdn = 17$, $p \text{ adj.} = .15$). Additionally, higher education was associated with higher COSIMO scores (low $m = 15.15 \pm 1.98$, medium $m = 16.06 \pm 2.18$, high $m = 16.52 \pm 1.83$, high>low $p \text{ adj.} = .001$), where *low* was defined as <11 years of education; *medium* as non-academic education ≥ 11 years, and *high* as having an academic degree. There was also a trend towards a cross-effect between age and gender, where female participants achieved higher scores than male participants at younger ages but had a stronger decline and eventually had reduced scores at higher ages.

Table 2. Descriptive table of sample size and demographic characteristics

		Total	Norms	TLE	ASD	ADHD	Other
N		2299	700	36	115	71	1377
Age m (SD)		39.80 (15.99)	45.78 (17.88)	41.58 (13.86)	37.44 (13.50)	33.79 (10.93)	37.22 (14.53)
Gender n (%)	F	744 (67.5%)	398 (69.7%)	17 (47.2%)	76 (74.5%)	26 (63.4%)	227 (64.3%)
	M	359 (32.5%)	173 (30.3%)	19 (52.8%)	26 (25.5%)	15 (36.6%)	126 (35.7%)
Education n (%)	Low	117 (10.2%)	27 (4.6%)	6 (16.7%)	18 (17.0%)	7 (17.1%)	59 (15.6%)
	Med	392 (34.2%)	125 (21.4%)	25 (69.4%)	37 (34.9%)	12 (29.3%)	193 (51.1%)
	High	636 (55.5%)	432 (74.0%)	5 (13.9%)	51 (48.1%)	22 (53.7%)	126 (33.3%)
Location n (%)	Online	789 (34.3%)	510 (72.9%)	0 (0.0%)	101 (87.8%)	37 (52.1%)	141 (10.2%)
	Onsite	1510 (65.7%)	190 (27.1%)	36 (100.0%)	14 (12.2%)	34 (47.9%)	1236 (89.8%)

Note. Gender: F = female, M = male; Education: Low = < 11 years of education, Med = >11 years of non-academic education, High = academic degree; Location: Online = test was completed as a volunteer without guidance online, Onsite: test was completed at a clinic in the presence of a professional.

Table 3. Descriptive statistics and sample sizes for included tests

	Total	Norms	TLE	ASD	ADHD	Other
COSIMO	15.56 ± 2.34 n = 2299	16.23 ± 1.93 n = 700	14.81 ± 1.85 n = 36	14.28 ± 2.81 n = 115	15.82 ± 2.24 n = 71	15.33 ± 2.41 n = 1377
MoCA	26.31 ± 3.33 n = 102	27.88 ± 1.96 n = 48	24.32 ± 3.01 n = 22	-	28.50 ± 0.71 n = 2	25.13 ± 4.12 n = 30
MASC	31.61 ± 7.22 n = 31	34.00 ± 7.09 n = 8	27.71 ± 5.06 n = 7	25.33 ± 11.72 n = 3	-	33.69 ± 6.17 n = 13
FPT	16.33 ± 2.99 n = 47	17.25 ± 2.19 n = 18	15.94 ± 2.31 n = 8	13.88 ± 6.70 n = 4	15.17 ± 0.29 n = 3	16.32 ± 2.93 n = 14
RMET	24.76 ± 5.27 n = 624	25.87 ± 4.66 n = 351	19.00 ± - n = 1	21.63 ± 6.14 n = 97	24.39 ± 6.55 n = 28	24.29 ± 4.85 n = 147

Note. Mean ± SD and n for each group per test. MASC = Movie for the Assessment of Social Cognition, RMET = Reading the Mind in the Eyes Test, FPT = Faux Pas Test, COSIMO = Cognition of Social Interaction in Movies.

Norm model

Across the full norm sample, we observed a near-normal distribution, although the variance in the scores was not stable across different ages, genders, and education levels. Therefore, we generated two multiple regression models for the norm model: one for the variance to model the distribution across the demographic variables and one for the predicted scores. The z-scores were then calculated by checking the distance of an individual's actual score from the predicted score from the mean model, divided by the predicted standard deviations from the distribution model. For a visualisation of the percentile spread across age and gender groups, see **Table 4**. The following norm models were compared:

- Age (in full years)
- Gender (male or female)
- Education (low = < 11 years, medium = non-academic education ≥ 11 years, high = academic degree)
- Location (online versus onsite)

Table 4. Percentile values were divided by age group and gender.

Gender	Female				Male			
Age	16-24	25-44	45-64	65+	16-24	25-44	45-64	65+
N	41	198	122	37	14	71	64	24
CER Score								
21			100					
20	99	98	99		100	99	98	
19	89	89	95	100	93	94	90	
18	76	73	83	96	86	81	74	98
17	59	53	61	84	79	64	53	83
16	37	31	38	68	54	44	36	62
15	20	15	23	55	29	24	26	40
14	12	6	13	38		13	16	19
13	7	3	6	18	18	7	8	
12	2	1	3	5	7	4	5	8
11						1	3	4
10			1				2	
9								

Note. CER Score = total correct points in COSIMO out of 21. Where parts of the range are missing, *n* and variance were insufficient to calculate the full range.

A full summary of the model validation statistics is presented in **Table 5**. The model including only age and education achieved the best calibration in all estimates (RMSE = 1.78, MAE = 1.43, $r^2 = 0.11$, $z\ m = 0.04$, $z\ SD = 0.98$, $<1SD = 15.8\%$, $<2SD = 1.75\%$). Following this result, we calculated z-scores for all participants, for whom we knew their age and education. After applying the norm model, only participants with known age and education were included ($N = 1145$, norms $n = 584$, TLE $n = 36$, ASD $n = 106$, ADHD $n = 31$, Other $n = 378$) for any calculations henceforth.

Table 5. Comparison of alternative norm models.

Model	RMSE	MAE	R^2	$z\ m$	$z\ SD$	$< -1SD\ (\%)$	$< -2SD\ (\%)$
Age + education	1.78	1.43	0.11	0.04	0.98	15.79	1.75
Age + gender + education	1.79	1.44	0.10	0.04	0.99	12.28	3.51
Age + gender + education + location	1.79	1.44	0.09	0.04	0.99	12.28	3.51
Age + location	1.82	1.48	0.07	0.05	0.98	14.04	1.75
Age + gender	1.82	1.49	0.06	0.05	0.99	14.04	1.75
Age	1.89	1.52	0.04	0.04	1.07	14.39	3.67

Note. Age was included on a continuous scale, whereas gender, education, and location (online or onsite) were considered at a categorical level.

Convergent and Divergent Validity of COSIMO

We compared the performance of COSIMO to three well-established social cognition tests (Movie for the Assessment of Social Cognition [MASC⁴¹], Faux Pas Test [FPT^{32,42}], and Reading the Mind in the Eyes Test [RMET⁴³]) as well as one measure of general cognitive impairment (Montreal Cognitive Assessment [MoCA^{44,45}]) in subsamples during the onsite assessments (see Table 3 for details on the subsample distribution). Using Kendall's τ to address the non-normal distribution in these tests, we found overall significant correlations between the variables. The age- and education-adjusted COSIMO score correlated most strongly with the MASC ($\tau = 0.45, p = .024, p \text{ adj.} = .024, n = 25$), followed by MoCA ($\tau = 0.43, p < .001, p \text{ adj.} < .001, n = 66$), RMET ($\tau = 0.27, p < .001, p \text{ adj.} < .001, n = 542$), and FPT ($\tau = 0.414, p = .023, p \text{ adj.} = .024, n = 30$). Because all the included tests were originally designed to detect impairment rather than investigate ability on the full spectrum, we reran the correlations with only the patient groups (see Table 3 for details on the samples) to investigate the effect in the sample with the highest score variance. When only patients were considered, the correlation with MoCA ($\tau = 0.27, p = .064, p \text{ adj.} = .064, n = 47$) decreased and was no longer significant, while all the other correlations increased, although only RMET ($\tau = 0.34, p < .001, p \text{ adj.} < .001, n = 217$) remained significant. The remaining sample sizes in FPT ($\tau = 0.50, p = .043, p \text{ adj.} = .061, n = 17$) and MASC ($\tau = 0.49, p = .046, p \text{ adj.} = .061, n = 17$) may not have had adequate power to reach significance after corrections for multiple comparisons.

Detecting impairment with COSIMO

Receiver Operating Characteristic Curves (ROCs, **figure 2**), using the adjusted z-scores from the model adjusted for age and education, revealed acceptable area under the curve (AUC) results for detecting individuals with social cognitive impairment in the FPT and MASC ($AUC = 0.72, \text{false } n = 48, \text{true } n = 5$) or RMET ($AUC = 0.74, \text{false } n = 536, \text{true } n = 55$), as well as for detecting ASD ($AUC = 0.74, \text{false } n = 584, \text{true } n = 106$), while the detection of TLE ($AUC = 0.68, \text{false } n = 584, \text{true } n = 36$) diagnosis was just below acceptable. Although above random chance, COSIMO did not adequately detect ADHD ($AUC = 0.59, \text{false } n = 584, \text{true } n = 41$).

The sensitivity and specificity values at the cutoff scores for $z \leq -0.5, -1.0$, and -1.5 are presented in **Table 6**. When using RMET score $< -2SD$ or ASD diagnosis as detection condition, sensitivity and specificity were well balanced at a $z \leq -1.0$ with a sensitivity of 51.9-60.0% and specificity of 76.9-84.2%. For the other two detection conditions (MASC/FPT $\leq 2^{\text{nd}}$ percentile), the most balanced sensitivity-to-specificity values were found already at the $z \leq -0.5$ cut-off, with sensitivity at 55.6-60.5% and specificity at 66.7-77.6%.

Density plots of the distribution of z-scores in the different samples revealed overlaps with the norm group in the range of 59-79%, where the smallest overlaps were found for ASD and TLE and the largest overlap for ADHD (**Figure 3**).

Table 6. Sensitivity and specificity of COSIMO at different cut-off values

Cut-off	$z \leq -0.5$		$z \leq -1.0$		$z \leq -1.5$		n +	n -
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.		
FPT/MASC PR<2	60.5%	66.7%	20.0%	85.4%	0.0%	93.8%	5	48
RMET < -2SD	69.1%	67.5%	60.0%	76.9%	43.6%	87.7%	55	536
ASD diagnosis	64.2%	77.6%	51.9%	84.2%	55.6%	77.6%	106	584
TLE diagnosis	55.6%	77.6%	36.1%	84.2%	13.9%	91.8%	36	584

Note. Sens. = sensitivity, spec. = specificity, FPT = Faux Pas Test, MASC = Movie for the Assessment of Social Cognition, RMET = Reading the Mind in the Eyes Test, ASD = Autism Spectrum Disorder, TLE = Temporal lobe epilepsy, n + = sample size with condition ("positive"), n - = sample size without condition ("negative").

Language effects in the German and English versions

An additional sample of $n = 404$ healthy participants completed COSIMO in English. To explore whether the scores were comparable and to investigate the effect of the four removed and two reversed items (described in the Methods section on the development of COSIMO), we generated box plots of the English and German responses before and after the changes. These results revealed that, while the English-speaking group achieved a slightly higher mean score on the original scoring, the medians of the groups were

equal after the adjustments (**Figure 4**). Despite this, a two-sided Mann-Whitney-Wilcoxon U-test with Holm-Bonferroni correction revealed a significant difference with a small effect size as measured with Hedge's g after the adjustments (DE $m = 16.22 \pm 1.92$, EN $m = 15.67 \pm 2.45$, $g = 0.26$ [0.14–0.39], $U = 3012$, $p < .001$).

Reliability

COSIMO did not reach satisfactory internal consistency (Cronbach's $\alpha = 0.46$; McDonald's $\Omega = 0.46$; $n = 2199$). We also ran a bootstrapped split-half reliability test using Spearman-Brown with 5000 splits, which yielded similar results ($m_{SB} = 0.45 \pm 0.03$; $n = 2199$). Upon investigating reliability at the item level using the leave-one-item-out method, we found no items with particular strengths or weaknesses regarding their importance to reliability (highest change in $\alpha = 0.02$).

We calculated retest reliability based on a small sample of healthy online participants who completed the test twice on the same day to investigate whether there was a potential learning effect. We found a medium-to-large, near-significant correlation between the first and second test completion ($r = 0.56$ [-0.015, 0.860], $p = 0.056$, $n = 13$). To investigate any bias caused by the learning effect, we calculated a proportional bias test. No bias was observed ($r = -0.12$, $p = .715$) between the first and second participation.

Discussion

To our knowledge, COSIMO is the first social cognitive screening tool to be completely implemented in an independent web application with dynamic norming. COSIMO is an economically feasible online tool for screening social cognition with high ecological validity. The results were automatically calculated and adjusted for age, gender, and education using the models from the current study, providing a plotted z-curve with the individual's position marked on the curve (for example, Fig. 5). Fallback models are implemented when less information is available. Free access to the test can be requested on <https://cosimo-project.com>.

COSIMO can provide clinicians with useful information regarding the risk of social cognitive impairment. The AUCs for detecting impaired performance in existing social cognition tasks were 0.72–0.74, which is considered an acceptable sensitivity-to-specificity trade-off. The ideal cut-offs for balanced sensitivity and specificity varied depending on the criterion but were higher than expected at either $z \leq -1.0$ or $z \leq -0.5$. This may be because of potential noise in the norm group due to the online testing format, for example, noise caused by screen size, distractions during test completion, or participants with undiagnosed conditions, as well as possible natural variation in this skill in the general population. In addition, sensitivity and specificity may be attenuated when criterion measures differ from the screening tool in terms of construct scope, task format, and the social-cognitive subfunctions they assess. Variability in classification accuracy across criterion measures may therefore partly reflect differences in modality, ecological structure and task demands.

We adequately detected ASD (AUC = 0.74) and TLE (AUC = 0.68) but not ADHD (AUC = 0.58). This is in line with previous research, where social cognitive impairment is part of the core cognitive profile for ASD and TLE, but not ADHD. As a psychometric screening tool, COSIMO is intended to identify participants who may have challenges in everyday interactions. No diagnosis can be based on COSIMO results alone, as social cognitive deficits are not specific to one diagnosis, and there is a large overlap between norm and patient groups (Fig. 3).

When considering both patients and norms together, we found significant small-to-medium correlations ($r = 0.25$ – 0.45) for all three social cognition tests compared to COSIMO. When only the patient groups were considered, the correlations increased slightly ($r = 0.33$ – 0.50), although only one of the tests retained enough power to reach significance after correcting for multiple comparisons. In contrast, the MoCA, which we used to test for general cognitive impairment, showed the opposite pattern. A large ceiling effect in the norm group facilitated a stronger correlation, and upon removing the norm group from the sample, the correlation with MoCA decreased from $r = 0.43$ to $r = 0.27$ and was just above the significance threshold at 95% confidence, although the remaining 48 participants provided sufficient power. These results indicate that in the target groups, COSIMO correlated more strongly with social cognition tests than with general cognition.

Conceptually, COSIMO can be placed between the three social cognition tests as a complex, dynamic, and economic CER test (Fig. 6). Compared to basic emotion recognition, COSIMO requires the inclusion of additional information to correctly interpret the emotional content, such as dyadic-interactive context and body language, and the inference of more complex affective mental

states, increasing its complexity compared to traditional emotion recognition tasks. COSIMO fits well within the concept of complex emotion recognition (CER).

The norm group, consisting of $n = 700$ participants, was divided into education and gender groups for the norm model, with age being modelled continuously. All models are available in the online version, but based on this study, we recommend using the model including age and education corrections. Individuals aged 20 and above with middle-to-high education levels were well represented. There were approximately twice as many female participants, although the male group remained representative of age and education. Future studies should include data for participants under 20 years of age, males, and those with less than 12 years of education. For edge groups, it is recommended to use the model only corrected for age, as this boasts the highest sample size and has acceptable estimation errors (Table 5). The dynamic norming approach allows for the continued updating of norms in a collaborative, open-science fashion.

To fully utilise the online format, we collected data via online recruitment, internal clinical evaluation, and aggregated external clinical data. The latter was anonymous for data privacy reasons; therefore, we had a large sample with unknown patients. We included these data in the study to understand the distribution of patient responses compared to the norm group. Internal clinical evaluations were used to collect comparative data on other social cognitive tests and the MoCA; however, owing to time constraints, we could only perform one or two additional tests per participant. We prioritised the MoCA and one other social cognition test, depending on the study phase. This led to correlational comparisons in a smaller subset of the sample (see Table 3 for the sample size per test). Therefore, the results of these tests cannot be generalised to the full sample but provide an indication of how COSIMO compares with these tests.

When comparing the English- and German-speaking norm groups, we found a potential language effect on the results. The adjustments we made to the German version caused the score in the English version to become lower than in the German group, while it was higher before these adjustments. This indicates that COSIMO must be validated and adjusted at the item level for each language. Because the items were originally developed using an English-speaking sample, the items and response options were based on English. Therefore, the norm models available for the English language are based on the original 25 items. In the online version, participants see all 25 items, regardless of language, to ensure future comparability between the language versions. Separate adjustments for each language can be added to the database when sufficient validation data are available.

The low estimates of reliability and internal consistency for COSIMO should be viewed considering the construct, task format and clinical context. COSIMO assesses performance across complex, dynamic, and naturalistic social-cognitive stimuli that vary in context, features, content, and response options, using a low number of stimuli for screening. Under such conditions, a lower internal consistency is expected⁴⁶ as item responses may reflect distinct processes rather than indicators of a single, narrow construct. The short test with few stimuli limits reliability estimates and makes them sensitive to item heterogeneity. For an effective screening tool, the key is not maximal internal consistency for a narrow subfunction but clinically meaningful classification accuracy and construct validity through verisimilitude. COSIMO showed promising discriminative performance with a satisfactory balance of sensitivity and specificity. Thus, the findings support its use as a brief tool for identifying individuals with potential impairments in social cognition.

The present findings support the use of COSIMO for research and for the descriptive assessment of CER performance relative to the available normative reference data. COSIMO is not intended or validated for diagnosis, differential diagnosis, clinical risk estimation, treatment planning, monitoring, or as a stand-alone indicator of social-cognitive impairment. Its automatically generated score should be interpreted as a psychometric performance estimate and, when used in applied settings, embedded in a broader professional assessment.

In conclusion, COSIMO is a brief psychometric screening tool for assessing CER. Owing to the automatic calculation of individual results based on the adjusted norms, little prior knowledge or instruction is necessary to administer the test, making it suitable for screening situations. Therefore, COSIMO represents an important step towards a more accessible assessment of social cognition.

Methods

Data collection/recruitment

Data were collected online and in-clinic, as described in our first publication on COSIMO². Online participants were recruited via forums, mailing lists, QR code flyers, personal contacts, and conferences to create a large base of normative data, with the additional recruitment of participants with ASD or ADHD. Located at the Swiss Epilepsy Centre, our lab collected data from patients with epilepsy during routine clinical testing and inpatient stays. All methods were approved by the ethics committee at the Philosophical Faculty at the University of Zurich, Switzerland, and performed in accordance with the relevant guidelines and regulations. A beta version of COSIMO was shared with other clinics, and aggregated data, including birth year and patient status, were used to study the effect of patient status on the COSIMO score. The participating clinics were located in Switzerland, Germany, and Austria. The beta app currently has approximately 200 users. The top ten users tested 50–100 participants each. For data privacy, we will not report the exact clinics or individual information.

Materials

COSIMO

Development

The conceptualisation of COSIMO was based on panel discussions of the diagnostic gap within social cognition. While considering the Movie for the Assessment of Social Cognition (MASC) as the gold standard for ecological validity, we aimed to develop a shorter test with less cognitive load and language dependency than the MASC.

The videos were collected by two paid students who scoured Netflix, Amazon Prime, and Disney + for short clips of four to eight seconds, featuring two characters whose faces were visible throughout. They collected 600 videos focusing on cultural, gender, ethnic, and age diversity using a guideline table. RJ checked all the videos for compatibility with the study format. Some videos were cropped or brightened, and some were returned to students for replacement if the criteria were unmet.

We explored methods for administering the test, considering PsychoPy and other services. To ensure easy and free access for clinicians, we decided to develop a stand-alone web application. Owing to Python's popularity in psychology research and web development, we based the application on it. Using Django, we integrated a PostgreSQL backend with an HTML5 front end, adding features with JavaScript and CSS. A colourful theme was created to make the test engaging for the participants.

We aimed to define emotional states for video characters using the RMET's well-established word list, which has been translated into many languages. We minimised the original 94-word list by removing uncommon and potentially hard-to-understand words to aid neurological patients and eliminate the need for a glossary. Through pilot studies, we reduced the list by removing synonyms and hard-to-translate words, focusing on German, English, Norwegian, and Spanish, with most of the data from German. The list was narrowed down to 13 positive, 13 negative, and 2 neutral states. The videos were rated as positive, negative, or neutral by three raters, and neutral or ambiguous items were excluded. The remaining videos were rated on Amazon Mechanical Turk, focusing on the emotional state of the marked character. Data were collected in stages, excluding videos with low inter-rater agreement to optimise item composition. The details of these studies will be published separately.

Based on the pilot studies, we selected 75 items with the highest interrater agreement on the emotion displayed, with at least 30% agreement on a single emotion as a cutoff. For each item, out of the total of 15 complex emotion words, four response options were generated using the top three options and one contrast word from the opposite valence list (e.g. for a negative state, a positive option was added to the list).

COSIMO version selection and item level adjustments

As the MTurk stage focused on the English test version, a panel study adjusted the German translations. Afterwards, the items were further developed using additional MTurk studies in English and the recruitment of German-speaking participants via e-mail and contacts. Each participant viewed 25 or 50 of 75 videos. This was decided to limit the time consumption per participant to 5–10 minutes so as to more easily be able to recruit participants more easily. We selected videos with good response spacing, requiring at least 50% to choose the best of the four options. The second and third most popular options were often similar; therefore, only one was required for differentiation. To minimise cognitive load, we reduced the options to three, selecting one incorrect option of the

same valence ("almost correct"), with an interrater agreement of 5–10%. The wrong valence option was retained to detect severe impairment or unlikely response patterns, indicating malcompliance. Thus, each item had three options: "correct" (50%+), "almost correct" (5–10%), and "wrong." After redefining the options, 50 items were split into two test versions with randomized item and response option order. These beta versions launched psychometric piloting and research in September 2022.

We analysed the two test versions based on response agreement in the German norm group and sensitivity to ASD and TLE. Version B outperformed version A in sensitivity. Some items appeared reversed or ambiguous in German, possibly because of translation artefacts. We decided to retire version A and focus on improving version B.

Before examining the overall test effects, we assessed the individual item performance to refine the sensitivity to ASD or TLE. Using Fisher's exact test, we calculated odds ratios and p-values for ASD and TLE participants' correct responses compared to the norm group. Six items had odds ratios below 1, indicating that patients more often responded correctly than the norm group. To check for translation effects, we switched the "correct" and "almost correct" options and reviewed the norm group response rates (item easiness). Three items improved in odds ratios ($OR = 1.37\text{--}2.03$, $p = 0.14\text{--}<0.001$) and item easiness ($0.56\text{--}0.77$), with two reaching significance. One item slightly improved in easiness (0.55) but had a weak odds ratio ($OR = 1.14$, $p = .52$). The other two items showed slight odds ratio improvement ($OR = 1.23\text{--}1.39$, $p = .12\text{--}.35$) but had strong norm group disagreement (easiness = $0.27\text{--}0.33$). We excluded the four items that did not clearly improve after switching response options and included the two with significant odds ratios post-adjustments.

Administration and scoring

COSIMO is an online application that can be used on any device browser. Onsite participants used iPads, whereas online participants used their own devices. The app provides instructions for participants that can be read independently. An example item was provided for practice. After the instructions, participants could repeat them, ask questions, or start the test. The first video started immediately, featuring two actors (Fig. 7A). The yellow frame indicated the actor to focus on (Fig. 7iii). Videos last 4–8 seconds. After the video, three response options (Fig. 7i) and a repeat button (Fig. 7ii) appeared below a still image of the video. The videos can be repeated once before the button disappears. Once a participant selects an option, the next item played (Fig. 7C). The items were presented in a random order. Participants earned one point for each correct answer, with a total score of 21 points.

Online version

Before starting the study, online participants were presented with a consent form and could continue only after providing their affirmative consent. They then completed a questionnaire with two control questions for comprehension and compliance, demographic questions, and enquiries about known psychiatric or neurological diagnoses. Next, the participants viewed COSIMO's instructions. The first page requested that the device be rotated sideways and the brightness be increased for optimal performance. An example video was provided, explaining the task (see online preview at <https://cosimo-project.com/cosimo/try>). Finally, the participants chose to continue, repeat the instructions, or contact us with any questions. If they continued, the test was initiated.

Additional tests and questionnaires

Depending on the study phase, different tests were administered in addition to COSIMO. We will examine COSIMO's construct validity by exploring its psychometric associations with the established tests in the subsamples that completed these measures. Each test is described in more detail below.

Movie for the Assessment of Social Cognition

During the first clinical validation stage, from September 2022 to March 2023, we administered the Movie for the Assessment of Social Cognition (MASC)^{16,49} to patients and controls. Developed in Germany two decades ago, the MASC remains a sensitive, ecologically valid, and cognitively complex social cognition test. A 15-minute video follows four acquaintances spending an evening together, with the test taking up to 45 minutes as it pauses 46 times for multiple-choice questions about the characters' emotions, thoughts, and intentions. Administered on a laptop or tablet using Microsoft PowerPoint, the MASC includes three subscales. We

used the total correct score, with a maximum of 46 points, as studying the tests at a subscale level was beyond the scope of this study.

Faux Pas Recognition Test

The Faux Pas Recognition Test (FPT)^{32,42} was conducted at the Swiss Epilepsy Centre during the second clinical validation phase from March 2023 to May 2024 and from February to May 2026. Short FPT versions have been established for clinical use. For this phase, we used a short version with five stories that had been previously used in our lab⁴⁷. Each story is read by the patient before the clinician asks about an awkward act (i.e. a social faux pas) by a character and if it was due to ignorance, what was inappropriate, and why (cognitive ToM), and how the character affected by the faux pas felt (affective ToM)⁴⁸. A control question per story ensured comprehension but did not count towards the score. Participants could score a maximum of four points per story, totalling 20 points.

Reading the Mind in the Eyes Test

The Reading the Mind in the Eyes Test (RMET)^{43,49} was adapted for online use in the COSIMO web application as a comparison test for an online population, available only to participants after June 2024. Despite criticism for its lack of ecological validity and cultural diversity^{6,50,51}, it remains a primary test in clinical practice and research^{1,5}, with less focus on its improved versions^{16,52}. The RMET includes 36 black-and-white images of pairs of Caucasian eyes expressing emotions. Each item had four response options, and participants received one point for each correct response⁴⁹. We used the German validation by Pfalz et al.⁵¹ to define the cutoff for impairment.

Montreal Cognitive Assessment

The Montreal Cognitive Assessment (MoCA)^{44,45} is a well-established, interculturally valid screening tool for general cognitive impairment. The test takes ten minutes to complete and covers multiple cognitive domains aggregated in a single score, with the highest possible score being 30. Throughout the study phase, volunteer participants and patients at the Swiss Epilepsy Centre completed the MoCA test alongside COSIMO as a control measure for general cognitive impairment.

Online demographic survey

Online participants completed a demographic survey, which evolved over time. Initially, it required age, education, and gender. Later, due to additional variables of interest when patient groups were recruited online, they included social well-being, support and burden, first language, and immigration background. The original survey asked about general mental fitness for control group inclusion or exclusion, while the later version added a specific diagnostic question. Participants could select or list diagnoses to identify as part of the norm or a specific patient group.

For this study, we focused on the minimal survey responses (age, gender, and education), as well as diagnosis where available, to maximise the sample size.

Data analysis

All analyses were performed using Python v. 3.9.5, using a Jupyter notebook.

Data Cleaning

The data were stored in an online database and extracted to CSV files using a Python script. Survey responses, patient information, and test results were aggregated into a single CSV file. Language models detected diagnoses and first languages from text fields. The variables were cleaned by removing duplicates and ensuring correct type storage. The categories were ordered for ordinal variables. The dataset was limited to German and English participants. Test attempts and data from external studies in other languages were excluded from the extraction. A separate dataset of individuals who took the same test twice was extracted for the retest reliability analysis. For the main analysis, the dataset included only German data from healthy individuals or patients aged ≥ 16 years, excluding those with invalid age or unclear group classification. Additionally, a dataset with German and English norm data was exported to compare the language versions.

A total of $N = 3565$ participants completed COSIMO in German ($n = 2824$) or English ($n = 741$) between September 2022 and May 2026, either as a volunteer (online or in-person) or as part of a routine neuropsychological assessment. From the German-speaking group, participants who were not clearly registered as either control or patient were excluded ($n = 525$). For the English-speaking group, only norm data were considered to compare the English and German norm data ($n = 404$). The inclusion criteria for the norm group were that participants should have explicitly reported having no neurological or psychiatric diagnoses. Participants in the diagnostic groups had to have received a psychiatric or neurological diagnosis from a professional. The “other” diagnostic category included aggregated results from diagnostic settings in which the participants had either identified themselves as patients or had been tested in a patient setting without explicitly reporting the diagnosis.

Data exploration and model definition

The resulting dataset is summarised in a demographic table (Table 2) and a results table (Table 3). We plotted the distribution of responses split by age, gender, and education and visually investigated these plots to identify areas where an effect might be expected to inform the definition of the model equation. The most evident differences were in high versus low education, as well as in age in the female group, although a weak age effect was also observed in the male group.

Norm models were created using combinations of age, gender, education, and online or offline participation to train the models using ordinary least squares regression to estimate the expected mean and standard deviation for each participant. The models were then investigated for residual error variance using the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) and variance contribution using R^2 . The model with the best overall model statistics (age + education) was selected and used to calculate the z-scores for each participant. Because education was not available for all participants, only a subsample of $n = 1145$ was included in the analyses from this point onwards.

Construct validity

To evaluate COSIMO’s construct validity and situate it within the broader nomological space of neuropsychological assessment, we examined its correlations with established measures of social cognition (MASC, FPT, and RMET) and the MoCA as an indicator of general cognitive impairment. Because COSIMO was the only normally distributed test, we used Kendall’s τ from the SciPy library, a robust non-parametric method which is less sensitive to ceiling effects or outliers, to correlate the raw scores of each test with the adjusted z-values of COSIMO. We corrected for multiple comparisons using the Benjamini-Hochberg method from the Statsmodels library to adjust the p-values. The ceiling effects in the norm group were partially so extreme (e.g. in MoCA) and the variance so low (e.g. in FPT) that the correlations reported were still driven by the abnormal distribution. Because the variance was larger and the scores were more well-distributed within the patient groups, we also investigated correlations in the patient group alone. A report of the isolated norm group correlations was omitted as the distribution and variance in score was inadequate for two of the tests and did not provide any additional information for the other tests, where the variance was more similar between the groups.

Receiver Operator Curves and Sensitivity

To investigate the sensitivity of COSIMO to social cognitive impairment, as indexed by traditional tests for social cognition, we generated ROC curves and calculated the area under the curve (AUC) for the following conditions: percentile range < 2 in MASC or FPT based on internal norms; 2 SD or more below the mean in RMET, based on the German norms by Pfalz et al.⁵¹; ASD diagnosis; or TLE diagnosis. As a control measure, we calculated the AUC of ADHD diagnosis to determine whether COSIMO would detect a diagnosis that is less frequently associated with social cognitive impairment^{1,36,53}. For the diagnostic detection criteria, we excluded patients with diagnoses different from the target conditions. In addition, we investigated possible cut-off scores and their sensitivity and specificity for discovering impairments in established social cognition tests

Comparison of German and English norms

To investigate the effects of the adjustments we made to the German version of COSIMO during development, we performed the same adjustments to the English norms (i.e. removing four items from the analysis and reversing the top two options of two items) to investigate the assumption that the adjustments needed to be language-specific. For comparison, we generated box plots of the English and German median scores before and after adjustments. We used a two-sided Mann-Whitney-Wilcoxon U-test with Holm-Bonferroni correction to investigate whether the difference between the English and German mean scores was significant after adjustments.

Reliability

We examined the reliability of COSIMO using Cronbach's α and McDonald's Ω across the full German-speaking sample. In addition, we ran a bootstrapped test-half reliability test with 5000 random item splits using Spearman Brown. To investigate whether the removal of any one item could improve reliability, we ran a leave-one-out analysis of Cronbach's α across all items. We also investigated same-day retest reliability in a small sample of $n = 13$ online participants using Pearson's r correlation analysis for the scores between the first and second test sessions. In addition, we generated Bland-Altman Plots to look for systematic bias, such as practice effects, in the sample. Given the small retest subsample and same-day interval, these analyses were interpreted as exploratory evidence of short-term repeatability rather than a robust estimate of temporal stability.

Declarations

Data availability

A minimal dataset of the data used in this study is available on GitHub: https://github.com/rejoha/cosimo_public.

Code availability

The web application is available on <https://cosimo-project.com>. An account can be requested for data collection and direct norm comparison. A simplified code used for the analysis is available on GitHub: https://github.com/rejoha/cosimo_public.

Ethical statement

This study involves human participants and was approved by the Ethics Committee of the Philosophical Faculty at the University of Zurich (Approval ID 22.9.8). Participants provided informed consent to participate in this study before taking part or consented to the use of the data collected during their clinical assessment at the Swiss Epilepsy Centre. Completely anonymous aggregated user data were included from other sources.

Author contributions

R.J. designed the test, developed the web application, oversaw the data acquisition, performed the analyses, and wrote the manuscript. M.E. assisted in data acquisition, concept development, and edited the manuscript. H.J. supervised the project, contributed to its conceptual development, and edited the manuscript. All authors read and approved the manuscript.

Acknowledgements

This work was supported by the Swiss Epilepsy Foundation (Schweizerische Epilepsiestiftung), Zurich, Switzerland. The foundation provides institutional (intramural) research funding as the governing body of our institution; no specific competitive grant number was assigned. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Competing Interests

All authors declare no financial or non-financial competing interests.

References

1. Cotter, J. et al. Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. *Neurosci. Biobehav. Rev.* **84**, 92–99 (2018).
2. Johannessen, R. et al. Complex emotion recognition in neurological and neurodevelopmental conditions. *Acta Psychol. (Amst)*. **267**, 107114 (2026).
3. Steiger, B. K. & Jokeit, H. Why epilepsy challenges social life. *Seizure* **44**, 194–198 (2017).
4. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5 (R))* (American Psychiatric Association Publishing, 2013).

5. Eicher, M. & Jokeit, H. Toward social neuropsychology of epilepsy: a meta-analysis on social cognition in epilepsy phenotypes and a critical narrative review on assessment methods. *Acta Epileptologica* **4**, (2022).
6. Ziermans, T. B., Hajdúk, M., Pinkham, A. E. & SIRS Social Cognition Research Harmonization Group. Call to action on social cognition measures in clinical research. *Schizophr Res. Cogn.* **43**, 100400 (2026).
7. Cavieres, A., Acuña, V., Arancibia, M. & Escobar, C. Advances in the ecological validity of research on social cognition in schizophrenia: A systematic review of the literature. *Schizophr Res. Cogn.* **39**, 100333 (2025).
8. Eicher, M., Johannessen, R. & Jokeit, H. Social neuropsychology of epilepsy in the digital age: A narrative review on challenges and opportunities. *Epilepsy Behav.* **166**, 110336 (2025).
9. Osborne-Crowley, K. Social cognition in the real world: Reconnecting the study of social cognition with social reality. *Rev. Gen. Psychol.* **24**, 144–158 (2020).
10. Johannessen, R. *Cultural Differences in Social Cognition: Adapting NEMO to a South African Population* (University of Zurich, 2020).
11. Cerami, C. et al. Understanding barriers and optimizing socio-cognitive assessment in the diagnosis of neurocognitive disorders. *J. Neuropsychol.* **19**, 603–618 (2025).
12. Froming, K. B. et al. *Comprehensive Affective Testing System. User's Manual* (Psychology Software Tools, Inc, 2006).
13. Ekman, P. Are There Basic Emotions? *Psychol. Rev.* **99**, 550–553 (1992).
14. Ekman, P., Friesen, W. V., O'Sullivan, M. & Chan, A. Et al. Universals and cultural differences in the judgments of facial expressions of emotion. *J. Pers. Soc. Psychol.* **53**, 712–717 (1987).
15. Sauter, D. A., Eisner, F., Ekman, P. & Scott, S. K. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2408–2412 (2010).
16. Golan, O., Baron-Cohen, S., Hill, J. J. & Golan, Y. The reading the mind in films task: complex emotion recognition in adults with and without autism spectrum conditions. *Soc. Neurosci.* **1**, 111–123 (2006).
17. Montag, C. et al. Different aspects of theory of mind in paranoid schizophrenia: evidence from a video-based assessment. *Psychiatry Res.* **186**, 203–209 (2011).
18. Abu-Akel, A. & Shamay-Tsoory, S. Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* **49**, 2971–2984 (2011).
19. Hassannataj Joloudari, J. et al. Complex emotion recognition system using basic emotions via facial expression, electroencephalogram, and electrocardiogram signals: a review. *Front. Psychol.* **17**, 1682883 (2026).
20. Msika, E. F., Despres, M., Piolino, P. & Narme, P. Dynamic and/or multimodal assessments for social cognition in neuropsychology: Results from a systematic literature review. *Clin. Neuropsychol.* **38**, 922–962 (2024).
21. Gao, S., Wang, X. & Su, Y. Examining whether adults with autism spectrum disorder encounter multiple problems in theory of mind: a study based on meta-analysis. *Psychon Bull. Rev.* **30**, 1740–1758 (2023).
22. Pagni, B. A., Walsh, M. J. M., Rogers, C. & Braden, B. B. Social cognition in autism spectrum disorder across the adult lifespan: Influence of age and sex on Reading the mind in the Eyes task in a cross-sectional sample. *Front. Integr. Neurosci.* **14**, 571408 (2020).
23. Lievore, R., Crisci, G. & Mammarella, I. C. Emotion recognition in Children and adolescents with ASD and ADHD: A systematic review. *Rev. J. Autism Dev. Disord.* **12**, 535–565 (2025).
24. World Health Organization. *Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders* (World Health Organization, 2024).
25. Steiger, B. K., Muller, A. M., Spirig, E., Toller, G. & Jokeit, H. Mesial temporal lobe epilepsy diminishes functional connectivity during emotion perception. *Epilepsy Res.* **134**, 33–40 (2017).
26. Bryant, K. L., Flattery, R. & Schurz, M. C. N. Cambridge University Press,. The role of the temporal lobe in human social cognition. in *Cooperation and Conflict* 104–133 (2021).
27. Wang, W. H. et al. Theory of mind and social functioning in patients with temporal lobe epilepsy. *Epilepsia* **56**, 1117–1123 (2015).
28. Edmonds, D. et al. The human social cognitive network contains multiple regions within the amygdala. *Sci. Adv.* **10**, eadp0453 (2024).

29. Cristinzio, C. & Vuilleumier, P. The Role of Amygdala in Emotional and Social Functions: Implications for Temporal Lobe Epilepsy. *Epileptologie* **24**, (2007).
30. Pehrs, C., Zaki, J., Taruffi, L., Kuchinke, L. & Koelsch, S. Hippocampal-temporopolar connectivity contributes to episodic simulation during social cognition. *Sci. Rep.* **8**, 9409 (2018).
31. Costa, M. et al. Aversive memory formation in humans involves an amygdala-hippocampus phase code. *Nat. Commun.* **13**, 6403 (2022).
32. Stone, V. E., Baron-Cohen, S. & Knight, R. T. Frontal Lobe Contributions to Theory of Mind. *J. Cogn. Neurosci.* **10**, 640–656 (1998).
33. Ainsworth, M. et al. Viewing ambiguous social interactions increases functional connectivity between frontal and temporal nodes of the social brain. *J. Neurosci.* **41**, 6070–6086 (2021).
34. Ziaei, M., Arnold, C., Thompson, K. & Reutens, D. C. Social cognition in temporal and frontal lobe epilepsy: Systematic review, meta-analysis, and clinical recommendations. *J. Int. Neuropsychol. Soc.* **29**, 205–229 (2023).
35. Borhani, K. & Nejati, V. Emotional face recognition in individuals with attention-deficit/hyperactivity disorder: a review article. *Dev. Neuropsychol.* **43**, 256–277 (2018).
36. Bora, E. & Pantelis, C. Meta-analysis of social cognition in attention-deficit/hyperactivity disorder (ADHD): comparison with healthy controls and autistic spectrum disorder. *Psychol. Med.* **46**, 699–716 (2016).
37. Filley, C. M. Social cognition and white matter: Connectivity and cooperation: Connectivity and cooperation. *Cogn. Behav. Neurol.* **33**, 67–75 (2020).
38. Pitcher, D. & Ungerleider, L. G. Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends Cogn. Sci.* **25**, 100–110 (2021).
39. Gernsbacher, M. A. Challenging the assumption that autistic people lack a theory of mind. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging.* **10**, 1001–1002 (2025).
40. 17 U.S. Code. § 107 - Limitations on exclusive rights: Fair use. *LII / Legal Inform. Institute*
<https://www.law.cornell.edu/uscode/text/17/107>
41. Dziobek, I. et al. Introducing MASC: A movie for the assessment of social cognition. *J. Autism Dev. Disord.* **36**, 623–636 (2006).
42. Baron-Cohen, S., O’riordan, M., Stone, V., Jones, R. & Plaisted, K. Recognition of Faux Pas by Normally Developing Children and Children with Asperger Syndrome or High-Functioning Autism. *J Autism Dev. Disord* **29**, (1999).
43. Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y. & Plumb, I. The Reading the Mind in the Eyes Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child. Psychol. Psychiatry.* **42**, 241–251 (2001).
44. Nasreddine, Z. S. et al. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).
45. Thomann, A. E. et al. The Montreal Cognitive Assessment: Normative data from a German-speaking cohort and comparison with international normative samples. *J. Alzheimers Dis.* **64**, 643–655 (2018).
46. Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behav. Res. Methods.* **50**, 1166–1186 (2018).
47. Bauer, J. et al. A comparative study of social cognition in epilepsy, brain injury, and Parkinson’s disease. *PsyCh J.* **12**, 443–451 (2023).
48. Stone, V. E. & Baron-Cohen, S. Faux Pas Recognition Test (Adult version). Preprint at (1998).
https://docs.autismresearchcentre.com/tests/FauxPas_Adult.pdf
49. Baron-Cohen, S., Jolliffe, T., Mortimore, C. & Robertson, M. Another advanced test of theory of mind: evidence from very high functioning adults with autism or asperger syndrome. *J. Child. Psychol. Psychiatry.* **38**, 813–822 (1997).
50. Pavlova, M. A. & Sokolov, A. A. Reading language of the eyes. *Neurosci. Biobehav Rev.* **140**, 104755 (2022).
51. Pfalz, M. C., Meyer, A. H., Opwis, K. & Martin-Soelch, C. The reading the mind in the eyes test: Test-retest reliability and preliminary psychometric properties of the German version. *Int. J. Adv. Psychol.* **2**, e1–e9 (2013).
52. Kim, H. A. et al. Multiracial Reading the Mind in the Eyes Test (MRMET): An inclusive version of an influential measure. *Behav. Res. Methods.* **56**, 5900–5917 (2024).
53. Morellini, L. et al. Social cognition in adult ADHD: A systematic review. *Front. Psychol.* **13**, 940445 (2022).

Figures

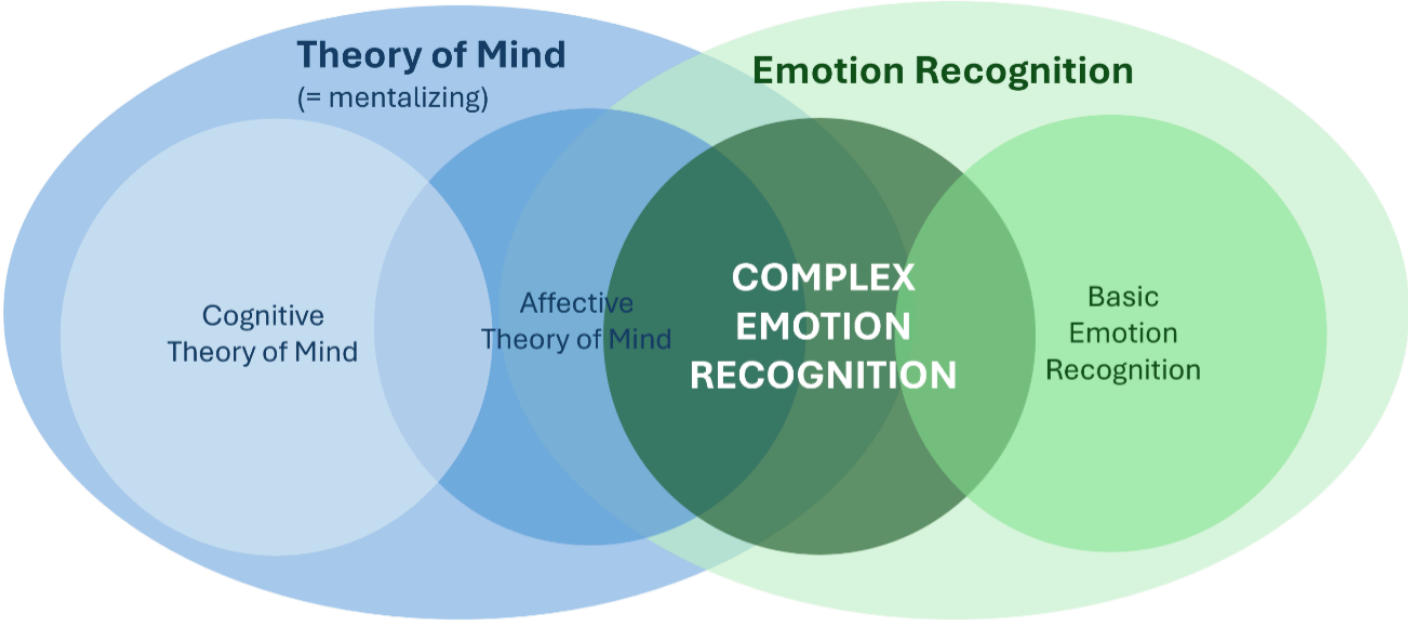


Figure 1

Venn diagram positioning complex emotion recognition within the framework of theory of mind and (basic) emotion recognition.

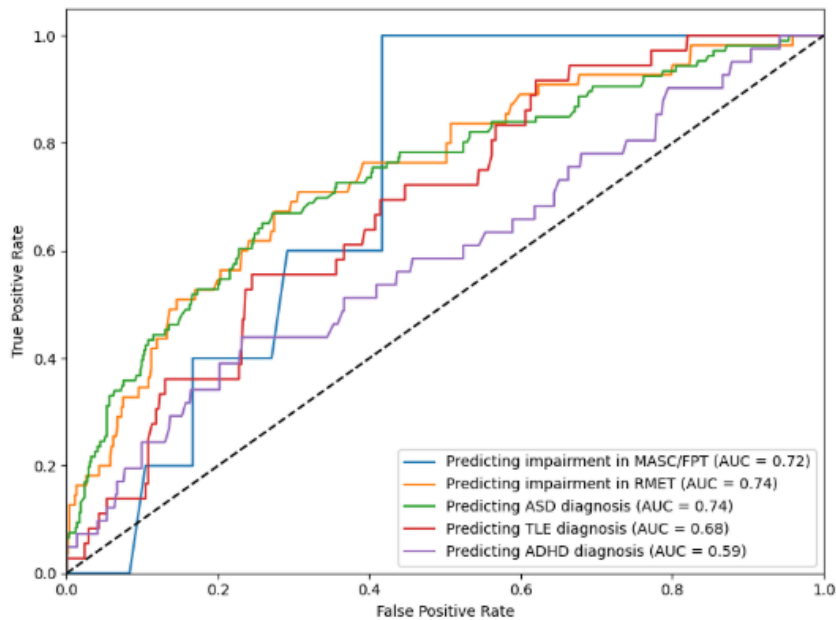


Figure 2

ROC plot for COSIMO under different prediction conditions.

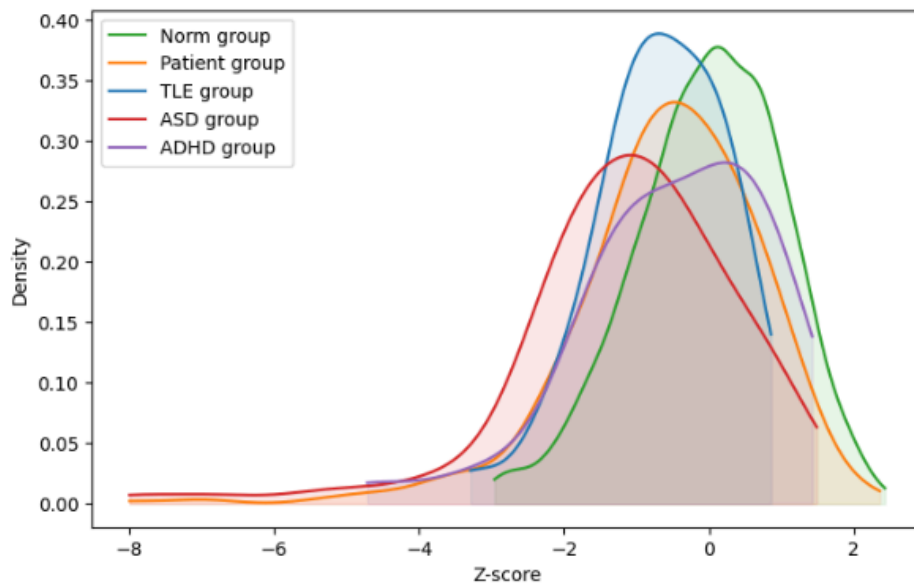


Figure 3

COSIMO Density plots for different population samples. Each curve represents the distribution of z-scores calculated using the age- and education-corrected norm model in a particular group.

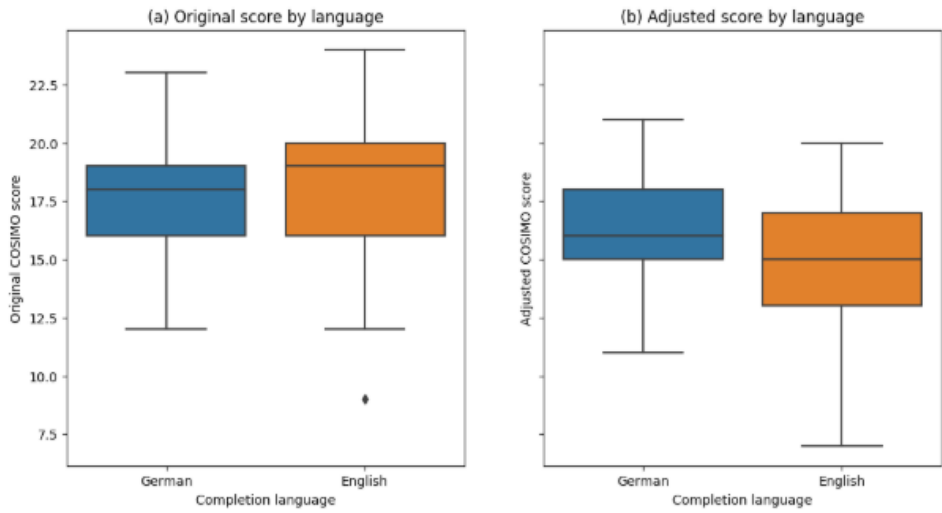


Figure 4

Comparison of scores for the English and German norm groups before and after the score adjustments.

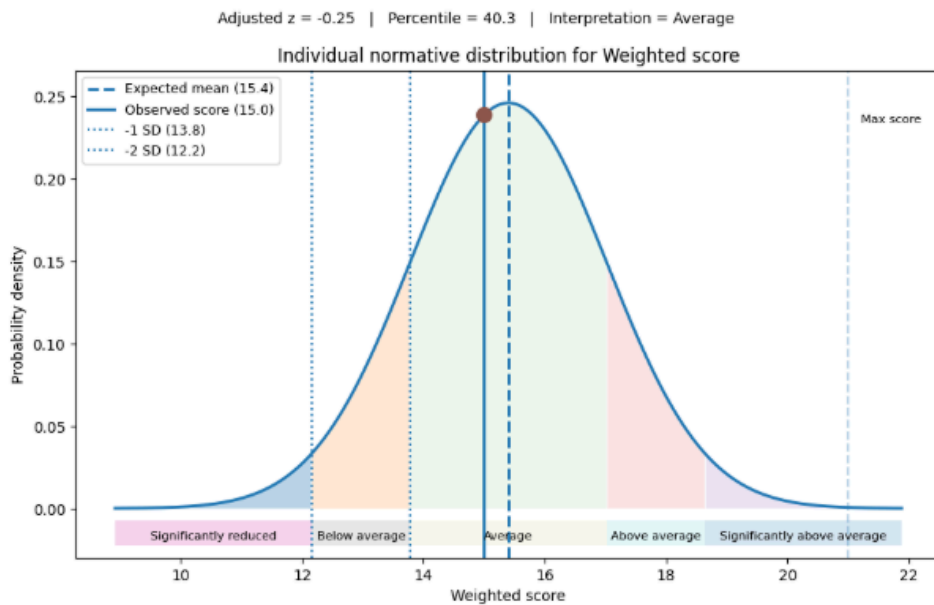


Figure 5

Example of a plotted score on a z-curve automatically output by the COSIMO application after completing the test. The red marker and solid blue line indicate the placement of the participant on the adjusted norm curve. Because of a slight skew, the maximum score is marked by a labelled dotted line. Cut-offs for 1 and 2 SD below and above the average are colour-coded in the area below the curve.

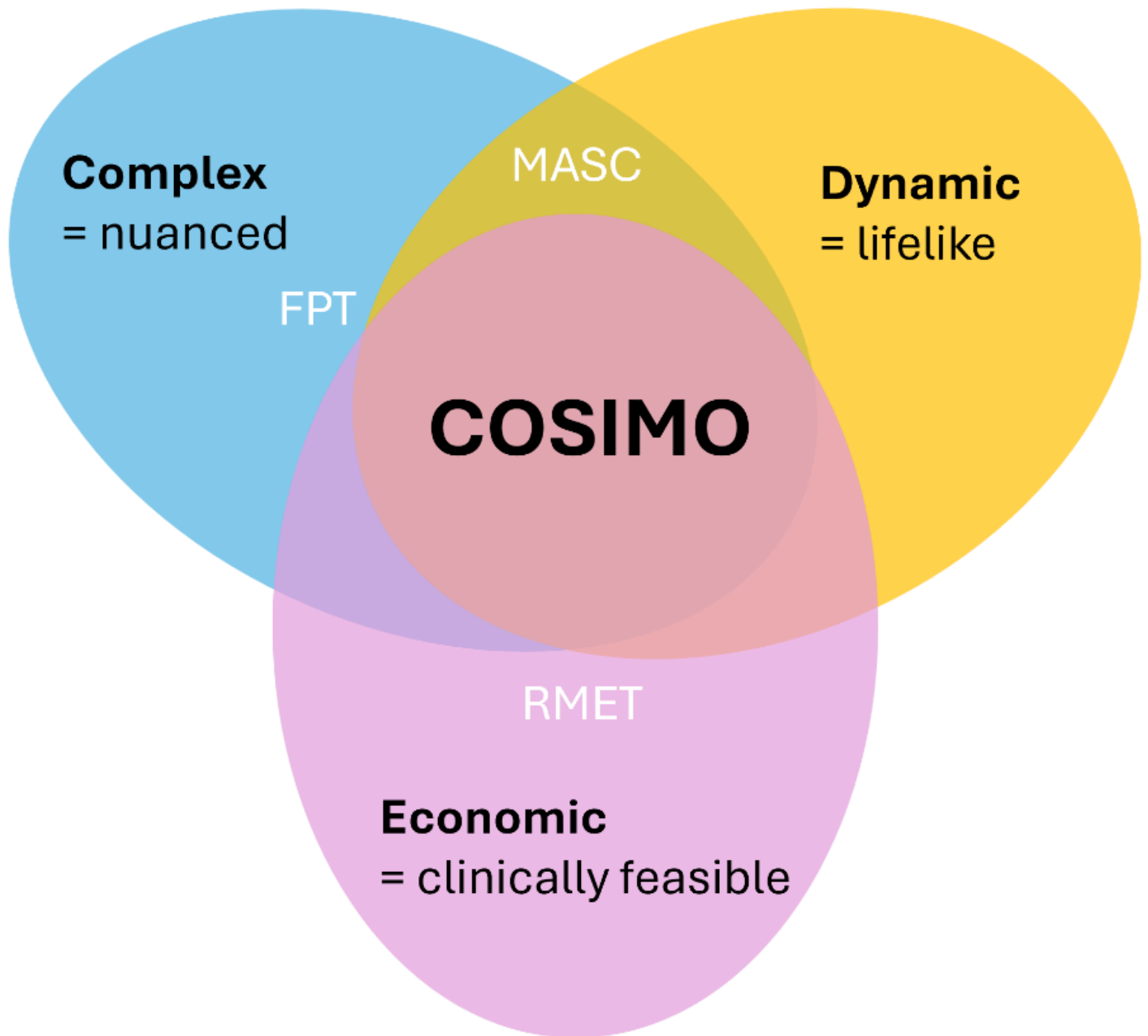


Figure 6

Figure placing COSIMO within a space of test features



Figure 7

Example items from COSIMO. This figure has also been published in a previous study by the same authors². (A) The first video was played automatically when the test began. (B) After the video, three options appeared at the bottom of a still image of the final frame. (C) As soon as a participant responds, the next video plays automatically until all 25 items have been responded to. (D) Participants could select one of three options, either by touch or by using a cursor (i). If necessary, they may watch the video again using the repeat button (ii). The yellow frame (iii) indicates whether the participants should focus on the left- or right-hand character.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [COSIMOPsychometricPaperSupplementaryMaterial.docx](#)