RESEARCH ARTICLE

Validity and reliability of the Frith-Happé Animation Test in a Turkish sample

Pelinsu Dilay Deniz[1] [ID], Ceren Hidiroglu Ongun[2] [ID]

[1]Canakkale Onsekiz Mart University, Faculty of Humanities and Social Sciences, Department of Psychology, Canakkale, Turkiye
[2]Dokuz Eylul University, Faculty of Letters, Department of Psychology, Izmir, Turkiye

ABSTRACT

Objective: The lack of assessments for social cognitive functions in Turkiye highlights the need for valid and reliable measurement tools in this field. This study aims to examine the reliability and validity of the Frith-Happé Animation Test (AT) in a Turkish sample. Additionally, it investigates the impact of demographic variables such as age, gender, and education level on AT performance, and explores differences in social cognitive functions between individuals with Autism Spectrum Disorder (ASD) and a healthy control group using the AT.

Method: The study included 267 healthy adults (145 females, 122 males) aged 18-45, along with 20 individuals diagnosed with ASD (four females, 16 males) aged 18-39. Participants were categorized by gender, age group (18-25, 26-35, 36-45 years), and educational attainment (lower: ≤12 years; higher: >12 years). The AT and the Dokuz Eylul Theory of Mind Scale (DEToMS) were administered.

Results: The AT demonstrated high internal consistency, with Cronbach's alpha coefficients of 0.673 for intentionality, 0.679 for appropriateness, 0.799 for certainty, and 0.906 for length. Test-retest analysis showed high stability in intentionality scores overall (r=0.835) and across different animation types. Criterion validity was moderate, with positive correlations between DEToMS total scores and intentionality scores (overall r=0.443; goal-directed r=0.368; theory of mind [ToM] r=0.437). Additionally, healthy individuals demonstrated better AT performance than those with ASD.

Conclusion: This study demonstrates that the Frith-Happé Animation Test is a valid and reliable measurement tool in a Turkish sample. In this context, it serves as an effective instrument for assessing social cognition and holds significant potential to contribute to future clinical assessments and scientific research.

Keywords: Autism spectrum disorder, Frith-Happé animation test, reliability, social cognition, validity

INTRODUCTION

Social cognition (SC) is the ability to perceive and interpret others' intentions, behaviors, and emotions, leading to appropriate responses (1). Key components of SC include emotion processing, social perception, attributional styles, and theory of mind (ToM). Theory of mind, the most extensively studied aspect, is divided into cognitive ToM, which involves inferring others' thoughts and beliefs, and emotional ToM, which focuses on understanding emotions (2). Additionally, ToM can be classified as explicit, involving conscious effort, or implicit, which occurs automatically (3).

Impairments in ToM functions are commonly observed in various psychiatric, neurological, and developmental conditions (4). These impairments may lead to significant difficulties in social functioning by negatively affecting an individual's ability to understand the emotions and thoughts of others, accurately interpret social cues, and respond appropriately in social situations (5). Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterized by difficulties in social communication, repetitive behaviors, and impairments in social cognitive functions, which are central to the disorder (6). Social cognitive deficits in individuals with ASD are thought to primarily stem from impairments in ToM (7). These deficits have been reported to persist into adulthood, not only during childhood (8).

Currently, social cognitive functions are assessed for various purposes, including understanding their nature in healthy individuals, examining impairments in clinical populations, conducting scientific research, and performing neuropsychological evaluations. Compared to the international literature, Turkiye lacks sufficient studies and measurement tools for assessing social cognitive functions. Most research in this area has focused on adapting existing international measurement tools for use with Turkish samples. Validity and reliability studies have been conducted for tests such as the Reading the Mind in the Eyes Test (9), the Faux Pas Recognition Test (10), and the Edinburgh Social Cognition Test (ESCoT) (11). Nevertheless, the Dokuz Eylul Theory of Mind Scale (DEToMS), a measurement tool developed in Turkiye, significantly contributes to the field by evaluating various components of ToM (12). In Turkiye, the most commonly administered tests typically focus on evaluating human characteristics such as facial expressions, language usage, and social cues. However, social cognitive functions can also be assessed through abstract scenarios involving the movement of various shapes (13). The Frith-Happé Animation Test (AT), widely used internationally, provides a new approach to assessing SC. It emphasizes the ability to attribute mental states and emotions to the movements of simple geometric shapes, rather than relying on human features (14). The AT evaluates participants' interpretations of animations, ranging from simple random movements to those suggesting complex mental states (15). There is a significant need in Turkiye for assessment tools that explore social cognitive functions from diverse perspectives and support a broader evaluation of these abilities.

This study aimed to assess the validity and reliability of the Frith-Happé Animation Test within a Turkish sample, as well as to determine its suitability for scientific research and clinical evaluation. Additionally, the study sought to investigate the impact of demographic variables, such as age, gender, and education level, on AT performance among healthy individuals. Finally, the study aimed to explore the ability of the AT to differentiate the social cognitive functions of individuals diagnosed with ASD from those of healthy counterparts. The findings are expected to support the wider dissemination of the test in Turkiye, thereby providing a strong foundation for cross-cultural studies on SC.

METHODS

Study Design and Setting
This study, designed to evaluate the validity and reliability of the AT in a Turkish sample, was conducted between October 2023 and April 2024. Permission to use the test was obtained from Dr. Sarah White of the Institute of Cognitive Neuroscience at University College London. Ethical approval was granted by the Dokuz Eylul University Social and Human Sciences Scientific Research and Publication Ethics Committee (E-87347630-659-553803). Informed written consent was obtained from all participants.

The study examined the effects of three independent variables on AT performance: gender (female, male), educational attainment (lower, higher), and age (18-25, 26-35, 36-45 years). The dependent variables were AT scores related to intentionality, appropriateness, certainty, and length.

Participants
A total of 267 healthy adults participated in the study, comprising 145 females (54.3%) and 122 males (45.7%). Participants were aged between 18 and 45 years, with a mean age of 29.88 years (standard deviation [SD]=9.71), and none had any psychiatric or neurological disorders. The age variable was categorized into three groups: 18-25 years, 26-35 years, and 36-45 years. Education levels were classified as lower educational attainment (≤12 years) and higher educational attainment (>12 years). Demographic information for the healthy participants is presented in Table 1.

The clinical sample consisted of 20 adults diagnosed with ASD (four females, 20%; 16 males, 80%), aged between 18 and 39 years (mean

Table 1: Demographic characteristics of healthy participants by age group and educational level

Educational attainment	Age group		
	18-25	26-35	36-45
Lower			
N	54	40	40
Gender (M/F)	22/32	20/20	20/20
Age	18.69±0.72	30.53±3.12	42.50±2.68
Higher			
N	51	41	41
Gender (M/F)	20/31	20/21	20/21
Age	21.16±1.19	29.32±3.29	42.27±2.66

M: Male; F: Female, Lower Educational Attainment (≤12 years); Higher Educational Attainment (>12 years).

age=20.75, SD=4.74). Following the release of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), the diagnostic criteria for autism spectrum disorders were revised, leading to the consolidation of all diagnoses under the umbrella term ASD (6). Accordingly, all participants in this study were assessed based on the DSM-5 criteria and treated as a single diagnostic group. Inclusion criteria required that participants have no additional psychiatric diagnoses.

Individuals with ASD were voluntarily recruited from those receiving treatment at Çanakkale Onsekiz Mart University, Faculty of Medicine, Department of Mental Health and Diseases, as well as from rehabilitation centers. Eligibility and diagnostic consistency were confirmed in collaboration with mental health professionals, including psychiatrists and psychologists from the respective institutions. Participants' medical histories were thoroughly reviewed, and individuals with co-occurring psychiatric or neurological disorders or intellectual disabilities were excluded. Only those who met all inclusion criteria and provided informed consent were included in the study. Additionally, participants with ASD were matched with a control group of 20 healthy individuals based on demographic characteristics such as age, gender, and educational attainment.

All participants were included in the study if they met the following criteria: no history of brain surgery, head trauma, intellectual disability, or alcohol/substance abuse within the past year, and no visual or auditory impairments that would interfere with test administration.

Measures

Demographic Information Form

This form collects participant information including age, gender, education level, employment status, occupation, psychiatric and neurological history, substance abuse history, and history of head trauma or surgery.

Clinical Information Form

Designed for participants diagnosed with ASD, this form includes questions about age at diagnosis, medication use, comorbid psychiatric or neurological diagnoses, and relevant family medical history.

Frith-Happé Animation Test (AT)

Developed by Abell et al. (16), the AT assesses ToM skills and consists of two exercises and 12 test animations. The silent animations feature scenes where a large red triangle and a small blue triangle move in different ways. The AT includes three conditions: random, goal-directed, and ToM-related movements, with four animations per condition. In the random condition, the triangles move independently and aimlessly. In the goal-directed condition, the triangles engage in physical interactions toward a specific objective. In the ToM condition, the triangles display movements that suggest social interactions such as teasing, persuasion, or surprise, allowing individuals to infer mental states. After each animation, the researcher recorded participants' responses to the question, "What happened in this animation?" The AT is evaluated based on the criteria of intentionality (0-5 points), appropriateness (0-3 points), certainty (0-3 points) and length (0-4 points). Intentionality scores reflect the extent to which participants attribute mental states and intentions to the triangles. Appropriateness scores measure how well the response aligns with the animation content. Certainty scores indicate the clarity and confidence of the participant's interpretation. Length scores assess how detailed the responses were. In addition to this subjective scoring method (14), which allows participants to provide open-ended responses, there is also an objective scoring method available (17).

Dokuz Eylul Theory of Mind Scale (DEToMS)

The DEToMS was developed to assess individuals' ability to recognize false beliefs, metaphors, irony, and faux pas. The test includes seven story-based tasks and two picture-based tasks. In the story tasks, participants answer questions related to the narrative. In the picture tasks, they are asked to choose the fourth

Table 2: Comparison of Frith-Happé Animation Test scores across age groups in healthy participants

	18-25 age Mean±SD	26-35 age Mean±SD	36-45 age Mean±SD	F (2, 264)	Post-Hoc
Intentionality	2.250±0.408	2.030±0.382	1.971±0.456	F=11.859 **p<0.001**	18-25>26-35 **(p=0.001)**
					18-25>36-45 **(p<0.001)**
					26-35=36-45 (p=0.643)
GD Intentionality	2.331±0.523	2.120±0.451	2.057±0.521	F=7.775 **p<0.001**	18-25>26-35 **(p=0.013)**
					18-25>36-45 **(p=0.001)**
					26-35=36-45 (p=0.698)
ToM Intentionality	3.600±0.703	3.199±0.681	2.991±0.808	F=16.884 **p<0.001**	18-25>26-35 **(p=0.001)**
					18-25>36-45 **(p<0.001)**
					26-35=36-45 (p=0.168)
ToM Appropriateness	1.900±0.504	1.757±0.425	1.720±0.465	F=3.888 **p=0.022**	18-25>26-35 **(p=0.001)**
					18-25>36-45 **(p=0.027)**
					26-35=36-45 (p=0.870)
Length	1.968±0.627	1.781±0.650	1.665±0.553	F=5.845 **p=0.003**	18-25>36-45 **(p=0.003)**
					18-25=26-35 (p=0.098)
					26-35=36-45 (p=0.450)

SD: Standard deviation; GD: Goal-directed; ToM: Theory of Mind. One-way analysis of variance (ANOVA) results are presented in the row under F (2, 264). A post hoc Tukey test was used to explore pairwise differences. p<0.05 was considered statistically significant (bold values).

picture from two options that best continues the sequence of three pictures. The total possible score on the test ranges from 0 to 16. The internal consistency coefficient of the scale was reported as 0.64, and the test-retest reliability was found to be 0.90 (12).

Procedures

The study began with adapting the test's administration and scoring instructions through a translation and back-translation process. Initially, the researcher translated the AT's administration and scoring materials into Turkish, incorporating feedback from subject matter experts. A language expert then translated the Turkish version back into English. This version was sent to Dr. Sarah White for comparison with the original. In this process, Dr. Sarah White's feedback was incorporated as needed, resulting in the final version of the test. Following this, a pilot study was conducted with a small group of 10 participants. Necessary adjustments were made based on verbal feedback from the participants regarding the overall application process and the clarity of the instructions.

The AT animations were presented in full-screen mode using Microsoft PowerPoint on a 15.6-inch screen with a 1080p resolution. Participants' responses were recorded with their consent in a quiet, face-to-face setting. To reduce fatigue and sequence bias, the order of the AT animations was randomized for each

participant. Half of the participants completed the AT before the DEToMS, while the other half began with the DEToMS. Scoring was carried out by two independent raters, including the researcher, yielding inter-rater reliability coefficients ranging from 0.72 to 0.93. The DEToMS test was administered in printed format and included both the story and picture tasks. Test-retest reliability of the AT was examined with a sample of 30 healthy participants, who completed the test again after a 21-day interval. The study concluded with a thank-you to all participants.

Statistical Analysis

The dataset was examined for outliers, normality, and homogeneity. An independent samples t-test was used to assess differences based on gender and education level, while a one-way Analysis of Variance (ANOVA) was conducted to explore differences among age groups. Tukey's analysis was employed to identify the specific groups contributing to these differences. To evaluate the reliability of the AT, test-retest reliability and Cronbach's alpha internal consistency coefficients were calculated. Additionally, the Mann-Whitney U test was used to analyze score differences between the ASD group and healthy individuals. Spearman correlation analysis was conducted to examine the relationship between AT scores and the DEToMS total score, as part of the criterion validity assessment.

Table 3: Comparison of Frith-Happé Animation Test scores by educational attainment in healthy participants

	Lower educational attainment Mean±SD	Higher educational attainment Mean±SD	t (265)	p	d
Intentionality	2.045±0.437	2.152±0.423	-2.030	**0.043**	-0.248
ToM Intentionality	3.140±0.796	3.448±0.720	-3.317	**<0.001**	-0.406
Appropriateness	1.868±0.363	1.999±0.355	-2.987	**0.003**	-0.366
RAND Appropriateness	1.860±0.583	2.045±0.558	-2.650	**0.009**	-0.324
Length	1.703±0.613	1.936±0.615	-3.097	**0.002**	-0.379

SD: Standard deviation; ToM: Theory of mind; Rand: Random. Lower educational attainment: ≤12 years; Higher educational attainment: >12 years. $p<0.05$ indicates statistical significance (bold values).

RESULTS

Investigation of AT Scores According to Demographic Variables

An independent samples t-test was conducted to assess whether AT scores differed by gender. The findings indicated no statistically significant differences between male and female participants in terms of general intentionality (p=0.784), appropriateness (p=0.412), certainty (p=0.882), and length (p=0.096) scores. Additionally, the analysis revealed no significant gender-based differences in intentionality scores across the random, goal-directed, and theory-of-mind conditions (p=0.249, p=0.860, p=0.471, respectively). Similarly, appropriateness scores for AT in the random, goal-directed, and ToM conditions did not differ significantly between genders (p=0.789, p=0.361, p=0.358, respectively).

Table 2 presents the findings of the ANOVA conducted to determine whether AT scores among healthy participants varied by age group. The results showed that the general intentionality scores of participants in the 18-25 age group were significantly higher than those in the 26-35 and 36-45 age groups (p<0.001). Likewise, the 18-25 age group obtained higher intentionality scores in both the ToM (p<0.001) and goal-directed (p<0.001) conditions. However, comparisons between the 26-35 and 36-45 age groups were not statistically significant for general intentionality (p=0.643), ToM intentionality (p=0.168), or goal-directed intentionality scores (p=0.698). Intentionality scores for random animations did not differ significantly between age groups (p=0.508). In terms of appropriateness scores, only in the ToM condition did the 18-25 age group score higher than the other groups (p=0.022). The comparison between the 26-35 and 36-45 age groups was not statistically significant for appropriateness scores in the ToM condition (p=0.870). No significant differences were found between age groups for general appropriateness (p=0.286), goal-directed (p=0.093), or random (p=0.112) conditions. There were also no significant differences in certainty scores across age groups (p=0.905). Length scores showed that participants in the 18-25 age group provided longer explanations than those in the 36-45 age group (p=0.003). However, no statistically significant differences were observed between the 18-25 and 26-35 age groups (p=0.098), or between the 26-35 and 36-45 age groups (p=0.450).

An independent samples t-test was conducted to evaluate the effect of educational level on the AT scores of healthy participants. The results revealed that individuals with higher educational attainment had significantly higher intentionality scores in both the general intentionality condition (p=0.043) and the ToM condition (p<0.001) compared to those with lower educational attainment. Conversely, no significant differences were found in the random (p=0.304) and goal-directed (p=0.457) conditions based on education level. Furthermore, individuals with higher educational attainment demonstrated better appropriateness scores (p=0.003), including specifically in the random animation condition (p=0.009). Notably, appropriateness scores did not show significant differences based on educational level in the goal-directed (p=0.132), and ToM (p=0.108) conditions. Additionally, certainty scores were consistent across educational levels (p=0.273). The analysis also revealed that individuals with higher educational attainment provided significantly longer explanations than those with lower educational attainment (p=0.002). The related findings are presented in Table 3.

Table 4: Test-retest reliability and criterion validity of the Frith-Happé Animation Test based on the Dokuz Eylul Theory of Mind Scale

	Test-retest	DEToMS
Intentionality	0.835 **(p<0.001)**	0.443 **(p<0.001)**
GD Intentionality	0.756 **(p<0.001)**	0.368 **(p<0.001)**
RAND Intentionality	0.740 **(p<0.001)**	0.267 **(p<0.001)**
TOM Intentionality	0.862 **(p<0.001)**	0.437 **(p<0.001)**
Appropriateness	0.849 **(p<0.001)**	0.425 **(p<0.001)**
GD Appropriateness	0.981 **(p<0.001)**	0.395 **(p<0.001)**
RAND Appropriateness	0.756 **(p<0.001)**	0.423 **(p<0.001)**
TOM Appropriateness	0.846 **(p<0.001)**	0.332 **(p<0.001)**

DEToMS: Dokuz Eylul Theory of Mind Scale; GD: Goal-Directed; RAND: Random; TOM: Theory of mind. p<0.001 values are statistically significant (bold).

Table 5: Comparison of Frith-Happé Animation Test scores between participants with autism spectrum disorder and healthy controls

	HC		ASD		U	p
	Mean rank	Sum of ranks	Mean rank	Sum of ranks		
Intentionality	29.33	586.50	11.68	233.50	23.50	**<0.001**
GD Intentionality	27.78	555.50	13.23	264.50	54.50	**<0.001**
RAND Intentionality	21.85	437.00	19.15	383.00	173.00	**<0.001**
ToM Intentionality	29.88	597.50	11.13	222.50	12.50	**<0.001**
Appropriateness	25.48	509.50	15.53	310.50	100.50	**0.007**
GD Appropriateness	26.58	531.50	14.43	288.50	78.50	**<0.001**
RAND Appropriateness	16.18	323.50	24.83	496.50	113.50	**0.018**
ToM Appropriateness	28.58	571.50	12.43	248.50	38.50	**<0.001**

HC: Healthy control; ASD: Autism spectrum disorder; GD: Goal-directed; RAND: Random; ToM: Theory of mind. p<0.05 values are statistically significant (bold).

Reliability Analysis of the Frith-Happé Animation Test

Cronbach's alpha coefficients were used to assess the AT's internal consistency. The results showed good internal consistency for the following scores: intentionality (α=0.673), appropriateness (α=0.679), certainty (α=0.799), and length (α=0.906).

Test-retest measurements for all AT score types were assessed using Pearson correlation analysis. For the intentionality score, a strong positive correlation was found between the first and last measurements (r=0.835, p<0.001). Similarly, strong positive correlations were observed for goal-directed animations (r=0.756, p<0.001), random animations (r=0.740, p<0.001), and ToM animations (r=0.862, p<0.001). For the appropriateness score, there was also a strong positive correlation between the first and last measurements (r=0.849, p<0.001), as well as for goal-directed (r=0.981, p<0.001), random (r=0.756, p<0.001), and ToM animations (r=0.846, p<0.001). Additionally, strong positive correlations were found between the first and last measurements for certainty

(r=0.704, p<0.001) and length (r=0.869, p<0.001). The related findings are presented in Table 4.

Validity Analysis of the Frith-Happé Animation Test

Correlation coefficients between the AT intentionality and appropriateness scores obtained from the healthy sample and the DEToMS total score were calculated to assess criterion-related validity. Only intentionality and appropriateness scores were selected, as they are considered more relevant to ToM skills. The results indicated a positive and moderate correlation between the total DEToMS score and the intentionality score (r=0.443, p<0.001), as well as the intentionality scores for goal-directed animations (r=0.368, p<0.001) and ToM animations (r=0.437, p<0.001). A positive but weak correlation was found between the intentionality score for random animations and the total DEToMS score (r=0.267, p<0.001). Similarly, there was a positive and moderate correlation between the total DEToMS score and the AT appropriateness score (r=0.425, p<0.001), as well as the appropriateness scores for goal-directed (r=0.395, p<0.001), random (r=0.423,

p<0.001), and ToM animations (r=0.332, p<0.001). The related findings are presented in Table 4.

When healthy individuals and individuals diagnosed with ASD were compared using the Mann-Whitney U test, healthy individuals were found to score higher across all AT score types, including intentionality and appropriateness scores for random, goal-directed, and ToM animations (p<0.001). Individuals with ASD scored higher than healthy participants only in the appropriateness scores for random animations (p=0.018). Table 5 presents the differences in intentionality and appropriateness scores between the two groups.

DISCUSSION

The Frith-Happé Animation Test demonstrated moderate to high internal consistency, high test-retest reliability, and moderate criterion validity within the Turkish sample.

The intentionality and appropriateness scores of the AT showed moderate internal consistency, while the certainty scores demonstrated higher internal consistency. The length scores exhibited the highest level of internal consistency. Consistent with previous studies (18, 19) that reported internal consistency coefficients for the AT, it can be concluded that the AT is a moderately to highly reliable measurement tool within the Turkish context. Similarly, the moderate internal consistency observed in the intentionality and appropriateness dimensions suggests potential areas for further refinement.

The measurements taken at two different time points for all AT score types showed high consistency. The strongest consistency between the two measurements was observed in the length of participants' responses, while the lowest consistency was found in the certainty scores, which reflect the extent to which participants hesitate or pause in their responses. For intentionality scores, the strongest relationship was observed in the ToM animations, followed by goal-directed and random animations. Similarly, for appropriateness scores, the strongest correlation was found in the goal-directed animations, followed by ToM and then random animations. These results indicate that reliable outcomes were obtained for the intentionality and appropriateness dimensions of the AT. The findings are partially consistent with those of a previous study (19). However, in contrast to that study, intentionality scores in the Turkish sample also demonstrated high consistency over time.

When individuals diagnosed with ASD were compared to demographically matched healthy adults on their AT scores, healthy individuals were found to perform significantly better in AT than those with ASD. Previous studies (17, 20), particularly the initial study (16) that introduced the AT, similarly reported poorer performance among participants with ASD. This finding suggests that healthy individuals not only possess stronger ToM skills but also engage in mentalization even in animations that do not specifically require it, such as random and goal-directed scenarios. Additionally, healthy individuals demonstrated higher appropriateness scores in goal-directed and ToM conditions, whereas individuals with ASD showed higher appropriateness scores in random animations. Although this may initially appear counterintuitive, it likely reflects logical and meaningful patterns related to social cognitive processes. Specifically, it may be related to differences in cognitive processing styles commonly observed in individuals with ASD. Random situations in the AT are designed to depict scenarios without clear purpose or interaction. It is already well-established that individuals with ASD are less likely to attribute intentional or mental states to social stimuli (21). Accordingly, individuals with ASD may describe situations based on physical features, without assigning social or mental meaning. In other words, describing a random situation in the AT at a physical level, such as saying "just moving shapes" may align more closely with the expected appropriate explanation and, therefore, increase their appropriateness scores. In contrast, individuals with typical development tend to attribute more mental states, which may lead them to assign intentions, relationships, or emotions even to random animations. In such cases, excessive social interpretation may lower their appropriateness scores.

In the healthy sample, positive and moderately strong correlations were found between all AT score types and the total DEToMS score. The lowest correlation was observed between the DEToMS score and AT length scores, while the highest correlation was, as expected, between the DEToMS score and AT intentionality scores. The highest correlation was observed between the DEToMS score and AT intentionality scores in the ToM condition, followed by the goal-directed and random conditions. Regarding appropriateness scores, the strongest correlation was found between the DEToMS score and scores from random animations, followed by goal-directed and ToM conditions. Similarly, in the ASD group, the relationships between DEToMS and AT scores were also positive

and moderate. Our findings are similar to previous research showing that AT scores correlate weakly with false belief test scores (17) and moderately with Faux Pas task scores (22). To date, no studies have examined the relationship between DEToMS and AT scores in the literature. The modest correlation between the two tests can be attributed to several factors, including the distinct components of ToM assessed by each test, the different methodologies employed, and the fact that the tests evaluate ToM in different contexts. Specifically, DEToMS assesses multiple dimensions of ToM, such as the recognition of false beliefs and faux pas, while AT's intentionality scores focus on individuals' attributions of mental states. Additionally, DEToMS measures these skills using story- and picture-based tasks, whereas AT evaluates them through non-verbal cues involving animated characters devoid of human characteristics. Finally, differences in scoring methods between the assessments may also contribute to the weakened correlation. Additionally, while the stimuli in DEToMS are structured and context-based, those in the AT are abstract and dynamic. DEToMS was developed in Turkish to reflect Turkish culture and lifestyle, whereas the AT, although universally applicable, is more open to cultural interpretation differences. For these reasons, the lack of a high correlation between the two tests can be attributed to the fact that they assess ToM from different perspectives. This also highlights the importance of multi-method evaluations in SC research. In this context, it is recommended that the AT be used alongside other assessment tools in both research and clinical settings where a comprehensive evaluation of SC is required. In our study, the total DEToMS score, a composite measure reflecting overall ToM ability, was used to assess the validity of the AT. However, it is advisable for future studies to analyze the DEToMS subscale scores separately. This approach may help clarify the relationships between AT performance and the various subcomponents of DEToMS: representation, meta-representation, and empathy.

Several factors contribute to the AT's failure to achieve higher levels of validity and reliability. These include the open-ended nature of the animations and participant responses, which affects scoring; linguistic and cultural differences, which influence the application of scoring instructions; and the similarity of scores used in the AT (e.g., appropriateness scores range from 0 to 3), which can negatively affect validity and reliability calculations. The type of assessment used in this study for AT allowed participants to give free-form responses. While this approach offers several advantages, it can also result in subjective scoring. Although the high level of consistency observed between independent raters in our study indicates that subjectivity was prevented, it is recommended that future studies develop a more standardized system to reduce interpretive bias arising from linguistic and conceptual differences in participants' responses. The objective assessment method proposed by White et al. (17), or the combined use of both methods, may offer an alternative. Moreover, the use of natural language processing technologies could help reduce the researcher's workload and enhance measurement reliability by enabling automatic and consistent analysis of open-ended responses. Additionally, the limited scoring range in the AT may have caused participants' scores to cluster closely together, thereby reducing variability. Kaplan and Saccuzzo (23) emphasize that the correlation coefficient depends on variability, and that calculations can be negatively affected when there is a limited range in the measurements. Furthermore, when it is necessary to test the differences between groups, low score diversity may also limit the revealing of these differences. Therefore, the closeness of scores in AT may have reduced the sensitivity of the validity and reliability analyses, resulting in only moderate levels of validity and reliability.

This study also explored the influence of demographic variables on the performance of healthy adults in the AT. It was found that gender did not significantly affect any type of AT score, and that the SC functions assessed by the AT were comparable between males and females. This finding aligns with previous research (18, 19), which reported no gender differences in intentionality scores—a key measure of ToM skills as evaluated by the AT. Conversely, this result contrasts with other studies (24, 25) that suggested women outperform men in SC functions. This discrepancy may stem from the fact that those studies focused on specific aspects of ToM, such as understanding false beliefs, recognizing emotions, or detecting faux pas, which differ from the focus of the current study, as well as differences in the types of tests and methodologies employed. Given that gender differences in ToM performance can be influenced by various factors—including the type of emotion (positive or negative), the gender of the characters in the test, and the sensory modality (e.g., visual or auditory) (24) —the unique structure of the AT may not adequately capture these gender differences.

Young adults were found to have higher intentionality scores in both the general and goal-

directed, as well as the ToM conditions, compared to the other two age groups. This result suggests that young adults made more mental inferences than older participants and provided more appropriate and accurate explanations in the ToM condition. The extended responses given by young adults may reflect their greater motivation and interest in the task. In short, young adults perform better in situations that require mentalizing. This finding is consistent with previous studies (26, 27) examining the effect of age on ToM, as well as studies (22, 28) that measured ToM skills using the AT. Additionally, the current study found that young adults demonstrated a superior understanding of the scenarios depicted in the ToM condition by providing more appropriate and detailed responses, which is a novel contribution to the literature. It is known that the effect of age on ToM skills is observed regardless of task type, modality, or whether the function is emotional or cognitive (27). Compared to the youngest group in our study, the lower AT scores observed in the 26-35 and 36-45 age groups may indicate a decline in ToM skills with age. This trend could also be attributed to the abstract, context-independent, and spontaneous inference demands of the Frith-Happé test, as well as the possible use of less effective strategies by older age groups in such tasks.

In this study, individuals with higher educational attainment scored higher on AT intentionality scores for ToM animations and had higher appropriateness scores for random animations. They provided more mental state explanations and were better at identifying the underlying narratives of the random animations. They also gave longer and more detailed explanations of the animations overall. However, no differences were found between education groups in intentionality scores for goal-directed and random animations, not in appropriateness scores for goal-directed and ToM conditions. These findings are similar to those of a limited number of studies (29, 30) examining the effect of education level on SC functions. To date, no studies have directly investigated the impact of education on SC as assessed by the AT. Therefore, the consistency of these findings obtained through the AT with the general literature increases the external validity of our results.

Given the observed effects of age and education, researchers and clinicians should consider how test results vary across these variables to avoid misinterpretation. In this study, individuals aged 18 to 45 were included, and their education levels were categorized into two groups: lower and higher

educational attainment. While average AT scores for different age and education groups are presented in the tables, this information is insufficient. Future studies should consider including a broader age range, particularly older individuals. Additionally, it is recommended that normative values be established based on age, gender, and education level, taking into account differences in educational attainment.

This study has several strengths, including the matching of ASD and healthy control groups by age, education, and gender, as well as a relatively large sample size. However, certain limitations may affect the generalizability of the findings. Due to the difficulty of reaching individuals with only a primary school education, the AT data for low-educated participants was underrepresented. Additionally, since ASD is diagnosed approximately four times more frequently in males than in females (31), it was difficult to adequately include adult females over the age of 18 with an ASD diagnosis. As a result, the ASD group primarily consisted of adult males. The difficulty of accessing individuals with ASD who do not have comorbid conditions has generally led to limited sample sizes in ASD research. The limited sample size in this study, due to insufficient access to women with ASD and the exclusion of participants with comorbidities, limits the generalizability of the findings. Studies with more diverse samples may produce more accurate and generalizable results.

CONCLUSION

Social cognitive functions are crucial for human life, as they underpin social interaction and communication. Consequently, a variety of valid and reliable measurement tools are necessary to understand and evaluate the different dimensions of SC.

The results of the current study support the applicability of the AT for assessing SC in Turkiye, demonstrating moderate but acceptable validity and reliability. Moreover, the AT has the potential to distinguish between healthy individuals and those diagnosed with ASD in terms of ToM skills. Young adults with higher education levels exhibited stronger mentalizing abilities, as measured by the AT, while gender did not significantly impact AT performance.

Future research should aim to include a wider age range and more diverse clinical samples. Additionally, comparative studies are recommended to explore differences among individuals with low and middle levels of educational attainment.

Contribution Categories		Author Initials
Category 1	Concept/Design	P.D.D., C.H.O.
	Data acquisition	P.D.D.
	Data analysis/Interpretation	P.D.D., C.H.O.
Category 2	Drafting manuscript	P.D.D., C.H.O.
	Critical revision of manuscript	P.D.D., C.H.O.
Category 3	Final approval and accountability	P.D.D., C.H.O.
Other	Technical or material support	P.D.D., C.H.O.
	Supervision	C.H.O.

REFERENCES

1. Green MF, Penn DL, Bentall R, Carpenter WT, Gaebel W, Gur RC, et al. Social cognition in schizophrenia: An NIMH workshop on definitions, assessment, and research opportunities. Schizophr Bull 2008;34:1211-1220. [CrossRef]

2. Shamay-Tsoory SG. The neural bases for empathy. Neuroscientist 2011;17:18-24. [CrossRef]

3. Frith CD, Frith U. Implicit and explicit processes in social cognition. Neuron 2008;60:503-510. [CrossRef]

4. Cotter J, Granger K, Backx R, Hobbs M, Looi CY, Barnett JH. Social cognitive dysfunction as a clinical marker: A systematic review of meta-analyses across 30 clinical conditions. Neurosci Biobehav Rev 2018;84:92-99. [CrossRef]

5. Henry JD, von Hippel W, Molenberghs P, Lee T, Sachdev PS. Clinical assessment of social cognitive function in neurological disorders. Nat Rev Neurol 2016;12:28-39. [CrossRef]

6. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. Washington (DC): American Psychiatric Publishing; 2013. [CrossRef]

7. Frith U. Mind blindness and the brain in autism. Neuron 2001;32:969-979. [CrossRef]

8. Velikonja T, Fett AK, Velthorst E. Patterns of nonsocial and social cognitive functioning in adults with autism spectrum disorder: A systematic review and meta-analysis. JAMA Psychiatry 2019;76:135-151. [CrossRef]

9. Yıldırım EA, Kaşar M, Güdük M, Ateş E, Küçükparlak İ, Özalmete EO. Gözlerden Zihin Okuma Testi'nin Türkçe güvenirlik çalışması. Turk Psikiyatri Derg 2011;22:177-186.

10. Şandor S, İşcen P. Faux-Pas Recognition Test: A Turkish adaptation study and a proposal of a standardized short version. Appl Neuropsychol Adult 2023;30:34-42. [CrossRef]

11. Tanrıverdi O. Edinburgh sosyal biliş testi'nin (ESBİT) Türkçeye uyarlanması, geçerlilik ve güvenilirliğinin incelenmesi [master's thesis]. İstanbul: İstanbul Medipol Üniversitesi; 2022.

12. Değirmencioğlu B, Alptekin K, Akdede BB, Erdil N, Aktener A, Mantar A, et al. Şizofreni hastalarında Dokuz Eylül Zihin Kuramı Ölçeği'nin (DEZİKÖ) geçerlilik ve güvenilirlik çalışması. Turk Psikiyatri Derg 2018;29:193-201.

13. Heider F, Simmel M. An experimental study of apparent behavior. Am J Psychol 1944;57:243-259. [CrossRef]

14. Eddy CM. What do you have in mind? Measures to assess mental state reasoning in neuropsychiatric populations. Front Psychiatry 2019;10:425. [CrossRef]

15. Castelli F, Happé F, Frith U, Frith C. Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. Neuroimage 2000;12:314-325. [CrossRef]

16. Abell F, Happé F, Frith U. Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. Cogn Dev 2000;15:1-16. [CrossRef]

17. White SJ, Coniston D, Rogers R, Frith U. Developing the Frith-Happé animations: A quick and objective test of Theory of Mind for adults with autism. Autism Res 2011;4:149-154. [CrossRef]

18. Andersen NK, Rimvall MK, Jeppesen P, Bentz M, Jepsen JRM, Clemmensen L, et al. A psychometric investigation of the multiple-choice version of Animated Triangles Task to measure Theory of Mind in adolescence. PLoS One 2022;17:e0264319. [CrossRef]

19. Shahrivar Z, Tehrani-Doost M, Khorrami Banaraki A, Mohammadzadeh A. Normative data and reliability of the moving shapes paradigm. Iran J Psychiatry 2020;15:143-149. [CrossRef]

20. Livingston LA, Shah P, White SJ, Happé F. Further developing the Frith-Happé animations: A quicker, more objective, and web-based test of theory of mind for autistic and neurotypical adults. Autism Res 2021;14:1905-1912. [CrossRef]

21. Baron-Cohen S. Theory of mind and autism: A review. In: Glidden LM, editor. International review of research in mental retardation. San Diego: Academic Press; 2000. p. 169-184. [CrossRef]

22. Zhang X, Lecce S, Ceccato I, Cavallini E, Zhang L, Chen T. Plasticity in older adults' theory of mind performance: The impact of motivation. Aging Ment Health 2018;22:1592-1599. [CrossRef]

23. Kaplan RM, Saccuzzo DP. Psychological testing: Principles, applications, and issues. 9th ed. Boston (MA): Cengage Learning; 2017.

24. Thompson AE, Voyer D. Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. Cogn Emot 2014;28:1164-1195. [CrossRef]

25. Wacker R, Bölte S, Dziobek I. Women know better what other women think and feel: Gender effects on mindreading across the adult life span. Front Psychol 2017;8:1324. [CrossRef]

26. Bottiroli S, Cavallini E, Ceccato I, Vecchi T, Lecce S. Theory of mind in aging: Comparing cognitive and affective components in the faux pas test. Arch Gerontol Geriatr 2016;62:152-162. [CrossRef]

27. Henry JD, Phillips LH, Ruffman T, Bailey PE. A meta-analytic review of age differences in theory of mind. Psychol Aging 2013;28:826-839. [CrossRef]

28. Gigi A, Papirovitz M. Theory of mind in older adults: The nature of the decline. Appl Neuropsychol Adul 2025;32:92-97. [CrossRef]

29. Franco MG, Smith PK. Theory of mind, old age, and educational opportunity in Colombia. Int J Aging Hum Dev 2013;77:107-125. [CrossRef]

30. Giovagnoli AR. Theory of mind across lifespan from ages 16 to 81 years. Epilepsy Behav 2019;100:106349. [CrossRef]

31. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: A systematic review update. Autism Res 2022;15:778-790. [CrossRef]