Investigation of the reliability of the "Reading the Mind in the Eyes Test" in a Turkish population

Ejder Akgün YILDIRIM[1], Muzaffer KAŞAR[2], Mehmet GÜDÜK[3], Elif ATEŞ[2], İlker KÜÇÜKPARLAK[2], Erdem Onur ÖZALMETE[4]

SUMMARY

Objective: The "Reading the Mind in the Eyes Test"(RMET) is one of the most widely used tasks for examining the Theory of Mind. In this study, the goal was to determine the reliability of a Turkish version of the RMET and the distribution of the results in a healthy population.

Method: The original test is made up of 36 pictures of solely the eye region of different individuals. Subjects are asked to choose the option that best describes the mental state of individual in the picture. After 3 piloting studies each conducted with different sets of 8 people, 2 items (25 and 35) in which most of the subjects consistently replied to foil words, were excluded in final step. 117 healthy volunteers were given the test, and 70 of these individuals were re-tested two weeks later. The Bland and Altman method was used to examine test re-test reliability

Results: The mean value of the number of correct responses was 24.46 (SD=3.44) for the test and 24.13 (SD=4.36) for the re-test. No significant difference was shown between test and re-test results (p=0.815). After excluding items 19 and 21 which failed to show reliability, the mean correct response rates were 23.64 (SD=3.38) and 23.40 (SD=4.32), respectively. The educational level significantly affected the correct response rates.

Conclusion: The Turkish version of the RMET was found to be reliable in a healthy population. The educational level should be taken into account. The 32-item version of the Turkish RMET can be used to evaluate the social, cognitive and emotional processes in adults

Key Words: Theory of Mind, Reliability, Cognition, Emotion

INTRODUCTION

The "Reading the Mind in the Eyes Test" (RMET; Baron-Cohen et al., 1999, 2001) has often been used in studies investigating associations between social cognition and psychopathology (Lee et al. 2005, Murphy, 2006, Bora et al. 2006, Fertuck et al. 2009). The RMET is believed to tap into "mind reading" abilities, which is an important aspect of "Theory of Mind" (Baron-Cohen et al. 2001).

Theory of Mind (ToM) was first introduced by primatologists (Premack and Woodruff, 1978) who asked whether chimpanzees were able to attribute mental states to con-specifics. The term was later adopted by child psychologists who examined the ontogenetic development of this cognitive domain (Leslie, 1987). In general terms, theory of mind can be defined as the cognitive capacity to make inferences and/or to produce mental representations about the mental status of self and others. In the context of a primate group living with strong mutual dependency and complex interactions, it is the ability to identify others who cooperate, and more importantly who try to defect (Brüne and Brüne-Cohrs, 2006). A universal cheater detection mechanism in humans was shown experimentally (Cosmides 1989). ToM might be regarded as a product of selection due to social environmental pressures. ToM encompasses many of the characteristics that we associate with most aspects of "humanity" that are perceived positively such as

Received: 09.11.2010 - **Accepted:** 17.01.2011

Acknowledgement: We would like to thank Professor Simon Baron-Cohen and Dr. Sally Whellwright from the ARC for their endless support and Professor Martin Brüene for his recommendations and contributions to the text.

[1]Psych., [2]Psychiatry Resident, Bakirkoy Research and Training Hospital for Psychiatry, Neurology and Neurosurgery, BARILEM Evolutionary and Experimental Psychiatry Group, [3]Psych., Kars State Hospital, [4]Psych., Silvan State Hospital.
Ejder Akgün Yıldırım MD, e-mail: *ejderyildirim@yahoo.com*

empathy, compassion, and sympathy or negatively such as deceit, treachery and deception (Drubach 2007, Trivers, 1971). The human face has a unique function because it manifests people's social intentions and motivations (Schmidt and Cohn, 2001). Also, primate brain evolution is largely based on a vision dominated world which indicates the importance of gazes in mind reading (Dunbar 1998).

To date, various ToM tasks have been introduced since Wimmer and Perner's classical Sally and Anne test (1983). The commonly used are the Hinting Task (Corcoran et al. 1995), deception tests (Frith and Corcoran 1997), comprehension of metaphor and irony (Drury et al. 1998), picture sequencing (Langdon and Coltheart 1997), comic strips (Sarfati et al. 1997), Smarties Test (Perner 1987), Happe's strange stories (1994), and the Mental State Attribution Task (Brüne, 2005). However, most of these tests require complex mental processing and included control tasks solely to determine whether the subject has sufficiently acknowledged the story given. Also, some people who perform well in many ToM tasks might have a deficit in tests requiring an advanced mind reading ability, for example to put oneself into another's mind and "tune in" to his or her mental state, as is necessary in the case of the RMET (Baron-Cohen et al. 1997). The RMET differs from other ToM tasks in that it examines mental state decoding from facial features, whereas the other tasks are more based on mental state reasoning which does not necessarily require visual input (Bora et al. 2006).

The first version of the RMET published in 1997 included 25 items and two options for each item (one target, one foil). It was designed to measure social cognitive abilites in children with autism or Asperger syndrome. As the target words and foil words had opposite meanings and it was difficult to detect differences between adult clinical and non-clinical populations, the test was revised 2001 (Baron-Cohen et al. 2001). The revised version of the RMET included 36 items and four options for each item (one target, three foils). Also, some of the foil words were semantically closer to the target word which allowed for the detection of subtle differences in performance.

Initially, the RMET has been used as a part of evaluating social cognitive skills in autistic spectrum disorders, which are neurodevelopmental disorders characterized by impaired social interaction, communication and language as well as repetitive behaviours, restricted interests, lack of eye contact and joint attention. Studies have shown that patients suffering from brain lesions (Stone et al. 2003) or frontotemporal dementia (Torralva et al. 2006) have poorer performance on the RMET. The RMET has also been a point of interest in studies investigating various other psychopathological conditions, including psychopathy (Murphy 2006, Richell et al. 2003) and borderline personality disorders (Fertuck et al. 2009). Data from patients with psychotic disorders or mood disorders suggested that the RMET can be used as a valuable instrument providing numerical information that evaluates ToM abilities in these disorders (Bora et al. 2005, Irani et al. 2006, Kelemen et al. 2005). Some authors suggested that the RMET reflects emotional and affective processes of ToM (Bora et al. 2008). Patients regarded as clinically remitted or "well" might still be suffering from deficits in ToM abilities which might lead to poor social functioning and problems in interpersonal relationships (Bora et al. 2008). The RMET has also become a useful instrument for studies on the neurobiological substrates of social and emotional skills (Domes et al. 2007).

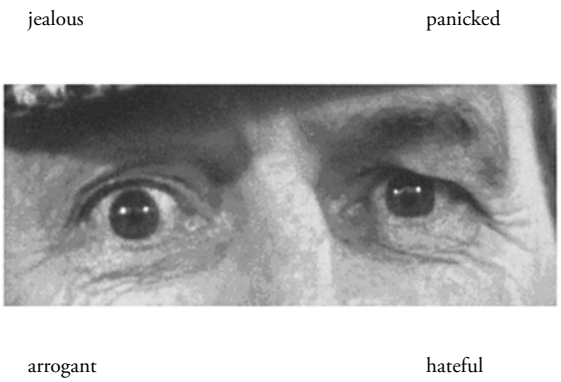
To date, many studies have used the RMET as a measure reflecting abilities such as facial emotion recognition, mental state decoding, or in a broad sense, ToM (Craig et al. 2004, Domes et al. 2007, Kelemen et al. 2004). As the test depends on complex mental processing including visual domains, emotion recognition and language related abilities, results may vary with regard to different test items, foil words, photograph properties etc. However, only one study (Hallerbäck et al. 2009) presented the full distribution of the results, hence it is difficult to compare possible response tendencies of various populations across different studies. Recently Erol *et al.* (2009) has determined the validity and reliability of two tests in the Turkish population called "Facial Emotion Identification" and "Facial Emotion Discrimination". These tests measure identification and discrimination of primary emotions (happiness, sadness, anger, fear, surprise, shame). The RMET has items addressing complicated emotions as well as thought processes (for example, irritated, despondent, flirtatious and puzzled were some of the options) and is designed to measure ToM performance along with emotion recognition.

In this study, the goal was to perform test-retest reliability of a Turkish version of the RMET and to determine its mean values and distribution of responses in a non-clinical Turkish population.

METHODS

Our own Turkish translation of the adult version of "Reading the Mind in the Eyes Test" (36 items) was used. The original test is composed of 37 photographs including one test item which is presented to familiarize the subjects with the test (Baron-Cohen et al. 2001). Each photograph has four options which include one target word and three foils (Picture 1). The subjects are asked to choose the option which best describes what the eyes in the photograph think or feel. Subjects are provided a glossary of 93 words describing the mental state terms on the cards and some semantically similar words. Definitions and example sentences were presented in the glossary. In the original study, subjects were encouraged to use the glossary because the mental state terms were not

commonly used in colloquial English. Every item has only one correct option. The total number of true responses were used to evaluate performance. A higher score indicates better social cognitive and ToM abilities.



Picture 1. Practice question in the Eyes Test

The Turkish adaptation of the RMET was courtesy of the Autism Research Centre (ARC/Autism Research Center – Cambridge, United Kingdom). The test was translated by two independent researchers who have an advanced level of English. Following the initial translations, a reverse translation was performed by a third researcher to determine the quality of translation. The authors discussed the appropriateness of each word as some of the words were rarely used in daily language and some of the words could only be expressed by using specific phrases in Turkish. Following suggestions by the ARC, we did pilot study on a group of 8 people (4 male, 4 female). After the pilot study, the items in which foil items were chosen by 4 or more subjects or items for which at least 2 subjects chose the same foil were re-evaluated in regards to translation and comprehensibility. We made subtle changes to terms which were often not comprehended and then a second pilot study was performed. Following the recommendations of the ARC, we conducted a third pilot study using the same criteria for false responses as described above and two items (Questions 25 and 35) were excluded as they were responded falsely by at least 5 of 8 subjects throughout three piloting. In our study, the version approved by the ARC which was comprised of 34 items plus a test item was used.

Participants

Ads for the study were used to recruit subjects and 130 participants were enrolled for this study, which was put on by the Bakirkoy Research and Training Hospital for Psychiatry, Neurology and Neurosurgery BARILEM Experimental and Evolutionary Psychiatry Group. All participants were interviewed using a short sociodemographic form prepared by the researchers. A detailed evalution for exclusion criteria were performed by psychiatrists. Our exclusion criteria were as follows: severe refractory deficit in the eyes, mental retardation, psychoactive substance abuse, taking psychotropic medication or having a history of psychotic disorders, recurrent mood disorders, any neurological disorders or a history of head trauma. We administered the first test to 122 participants. The responses from 5 subjects were excluded due to inaccurate filling of the test (i.e. either marking two options for an item or not responding to some of the items) and thus, data from 117 subjects was collected (72 female, 45 male). Two weeks later, 70 subjects underwent re-testing (39 female, 31 male). The study was approved by ethics commitee of the Bakirkoy Research and Training Hospital for Psychiatry, Neurology and Neurosurgery in June 2009.

Administration

When administering the test, each participant was given 35 cards including the test item (Figure 1) with the 93 word glossary. The participants were encouraged to use the glossary. The first page contained instructions and the second page contained the test item which was used for practice. The 34 item version was presented in the original format and order with one photograph and four options on each page. Response sheets were provided to record the options. The participants were not timed.

Statistical Analyses

We used the NCSS 2007&PASS 2008 Statistical Software (Utah, USA) program for statistical analysis. To evaluate normally distributed values, descriptive statistics (Mean, Standard deviation) and the Kolmogorov-Smirnov test was used. When test scores were not distributed normally, the Mann Whitney U test was used. Spearman coefficients were used for the evaluation of correlations between parameters. For reliability analysis of the RMET, we used the Bland Altman method (1986) and Kuder-Richardson 20. Correlations between test and re-test scores were evaluated with ICC (Intraclass Correlation Coefficient). The Kendall's Tau-b correlation was used to evaluate item based correlations based on true and false responses. The significance level was set at $p<0.05$.

RESULTS

Sociodemographics

The subjects' ages ranged from 19 to 57, with an average age of 33.33±8.95. Our sample had a female preponderance [72 female (61.5%), 45 male (38.5%)]. The level of education and distribution were as follows: 18.8% had primary education (n=22), 24.8% had high school education (n=29) and 56.4% had university education (n=66). 94.5% (n=110) of subjects were right-handed, 6% (n=7) were left-handed.

Distribution of responses

Overall, the mean value of the correct number of responses was 24.46 (SD=3.44) for the test and 24.13 (SD=4.36) for the re-test. Correct response rates were lower than 50% in items 19 and 21 (Table 1). When these two items (19 and 21) were excluded, the mean scores were 23.64 (SD=3.38) and 23.40 (SD=4.32), respectively. The Wilcoxon Signed Rank test did not show a significant difference between the correct response rates of test and re-test results (p=0.815).

Table 1. Distribution of responses for each item in Reading Mind in the Eyes Test – Turkish version (n=117)

Items	A (%)	B (%)	C (%)	**D (%)**
1	**65.0**	17.1	8.5	9.4
2	6.8	**83.8**	3.4	6.0
3	0.0	5.1	**85.5**	9.4
4	1.7	**80.3**	6.8	11.1
5	14.5	17.9	**67.5**	0.0
6	14.5	**71.8**	7.7	6.0
7	8.5	11.1	**56.4**	23.9
8	**82.9**	7.7	3.4	6.0
9	6.0	6.8	9.4	**77.8**
10	**70.1**	17.9	7.7	4.3
11	6.8	12.0	**77.8**	3.4
12	21.4	8.5	**63.2**	6.8
13	9.4	**76.9**	2.6	11.1
14	3.4	8.5	2.6	**85.5**
15	**81.2**	4.3	9.4	5.1
16	1.7	**84.6**	8.5	5.1
17	**56.4**	16.2	0.9	26.5
18	**87.2**	1.7	8.5	2.6
19	10.3	24.8	26.5	**38.5***
20	3.4	**78.6**	12.8	5.1
21	32.5	**42.7***	24.8	0.0
22	**86.3**	3.4	2.6	7.7
23	3.4	10.3	**60.7**	25.6
24	**69.2**	17.9	6.0	6.8
25	6.8	1.7	**70.9**	20.5
26	0.0	**62.4**	21.4	16.2
27	**89.7**	0.9	4.3	5.1
28	4.3	6.0	5.1	**84.6**
29	4.3	**76.9**	12.0	6.8
30	3.4	**70.9**	0.9	24.8
31	**72.6**	0.9	8.5	17.9
32	12.0	8.5	7.7	**71.8**
33	6.8	20.5	**54.7**	17.9
34	7.7	6.8	**65.8**	19.7

The correct response rates were significantly higher in female subjects (25.25±3.11) compared to males (23.20±3.5) (p<0,01). The educational level significantly affected the correct response rates (p<0.01). Subjects with primary education and subjects with high school education showed no significant difference (p>0.05). Correct response rates were significantly higher in subjects with university education compared to subjects with primary or high school education (p=0.001) (Table 2).

Table 2. Distribution of mean score according to education level and gender

Educational Level	Gender	n	Mean+SD	Median	*p*	
Primary		22	21.73±3.05	22	***0.001***	*Kruskal Wallis test*
High School		29	22.79±3.29	23		
University		66	26.11±2.62	26		
Primary	**Female**	13	22.54±3.38	22	***0.021****	*Mann Whitney U test*
	Male	9	20.56±2.18	21		
High School	**Female**	11	23.55±2.69	24	***0.403***	
	Male	18	22.33±3.61	22		
University	**Female**	48	26.38±2.48	26.5	***0.242***	
	Male	18	25.39±2.93	25.5		

In subjects with only primary education, the correct response rates were significantly higher for females however there was no significant difference between males and females either for subjects with high school education or subjects with university education (Table 2).

The correct response rates for the 32 item version were 73.87% overall and 81.59% for participants with university education. A comparison of correct response rates between the original study and the Swedish version were shown in Table 3.

Re-test

70 subjects completed the re-test. The correct response rates between test and re-test were evaluated using the Intraclass correlation coefficient (ICC) and a significant correlation was found [r=0.650; p<0.01 (95% CI: 0.490-0.768)]. Using Kendall's Tau-b correlation test to evaluate item based correlations according to true and false responses, the correlation coefficient for the 19th item was not significant (Table 4).

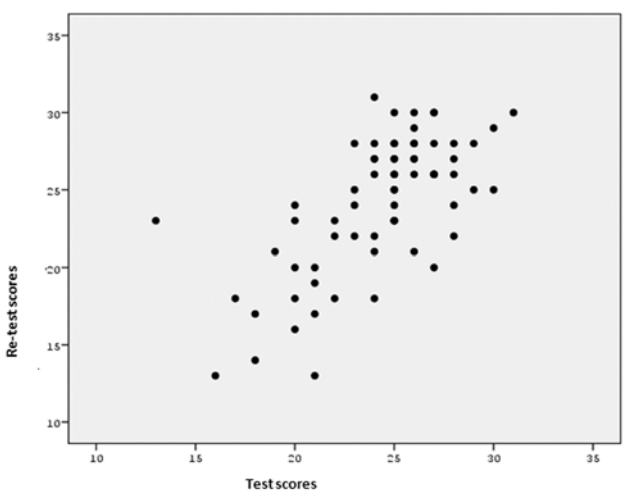


Figure 1. Correlation between test and re-test scores (correct responses)

Test re-test reliability was evaluated using the Bland Altman method (Bland&Altman, 1986). We observed that differences between test and re-test ranged between -10 and +8 and the

Table 3. Comparing correct response rates with original study and Sweedish Study

	Number of Items	Number of mean correct response		Response rates (number of correct response/ number of items)	
		All Participants	Participants with university education	All Participants	Participants with university education
Original Study *(Baron-Cohen et al 2001)*	36	26.2	28.0	72.77%	77.77%
Sweedish Study *(Hallerbäck et al 2009)*	24	_	18.9	_	78.75%
Turkish version	32	23.64	26.11	73.87%	81.59%

mean difference was 0.13 (SD=3.28). The confidence interval for the mean difference was 95% CI -0.65/0.91. When confidence intervals were calculated as X±2SD, the lower limit was -6.68 and the upper limit was +6.43 (Figure 1, Figure 2).

After the exclusion of two unreliable items (19 and 21) the differences ranged between -10 and +11 and the mean difference was 0.01 (SD=3.66). The confidence interval for the mean difference was 95% CI -0.86/0.88. When confidence intervals were calculated as X±2SD, the lower limit was -6.56 and the upper limit was +6.56 (Figure 2).

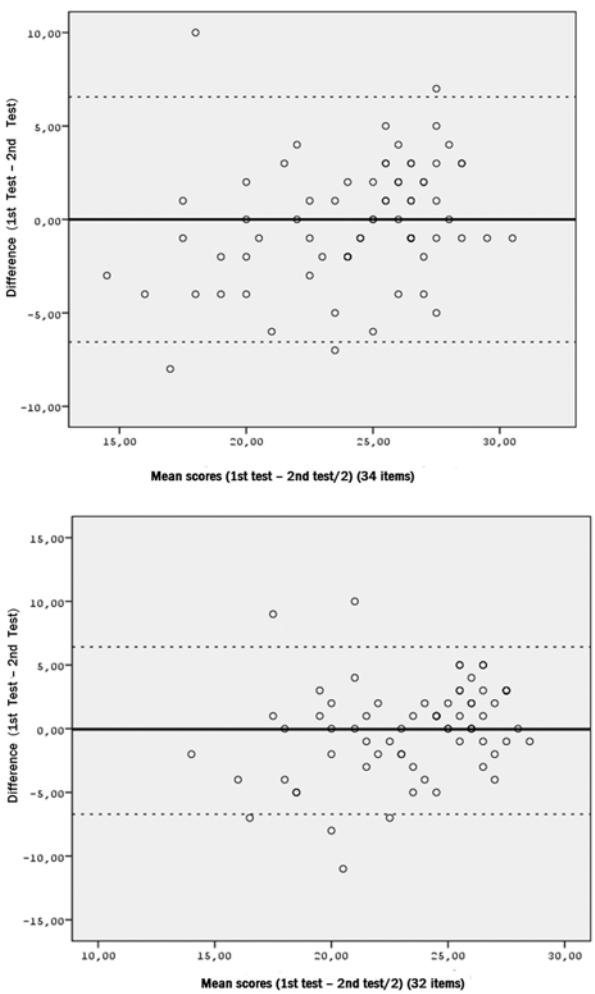


Figure 2. Distribution of two test results according to Bland Altman method (34 items and 32 items) (n=70).

Table 4. Analysis of test re-test responses for each item (n=70);

Items	One correct one wrong (%)	Both correct (%)	Both wrong (%)	Same response in test and re-test (%)	*Kendall's Taub*	
					r	p
1	15.7	58.6	25.7	74.4	0.648	*0.001*
2	18.6	74.3	7.1	77.1	0.330	*0.006*
3	11.4	81.4	7.2	85.7	0.538	*0.001*
4	18.6	74.4	10.0	79.9	0.339	*0.005*
5	23.0	57.1	19.9	74.2	0.495	*0.001*
6	24.3	61.4	14.3	68.5	0.376	*0.002*
7	31.3	42.9	25.8	62.9	0.371	*0.002*
8	11.4	77.1	11.5	85.7	0.598	*0.001*
9	27.2	62.9	9.9	68.6	0.247	*0.041*
10	35.8	48.6	15.6	61.5	0.306	*0.009*
11	15.8	67.1	17.1	81.5	0.581	*0.001*
12	30.0	57.1	12.9	62.8	0.271	*0.024*
13	20.0	67.1	12.9	75.7	0.433	*0.001*
14	19.9	71.4	8.7	78.5	0.361	*0.003*
15	24.4	70.0	5.6	71.4	0.297	*0.012*
16	17.2	80.0	2.8	82.8	0.253	*0.036*
17	34.3	40.0	25.7	61.4	0.304	*0.011*
18	18.5	75.7	5.8	80.0	0.349	*0.004*
19*	30.0	**20.0**	**50.0**	41.5	0.153	***0.203***
20	21.5	64.3	14.2	74.3	0.433	*0.001*
21*	34.2	**27.1**	**38.7**	57.1	0.306	*0.011*
22	15.7	77.1	7.2	82.8	0.385	*0.001*
23	24.3	44.3	31.4	62.8	0.532	*0.001*
24	34.4	51.4	14.2	62.8	0.305	*0.010*
25	27.2	58.6	14.2	68.6	0.325	*0.007*
26	35.7	47.1	17.2	57.1	0.295	*0.018*
27	25.7	70.0	4.3	70.0	0.302	*0.009*
28	24.4	72.9	2.7	72.9	0.510	*0.001*
29	27.2	62.9	9.9	65.8	0.257	*0.033*
30	29.9	57.1	13.0	67.0	0.257	*0.033*
31	31.4	52.9	15.7	67.2	0.271	*0.024*
32	35.8	51.4	12.8	58.6	0.264	*0.017*
33	24.4	41.4	34.2	58.5	0.515	*0.001*
34	27.2	50.0	22.8	65.7	0.414	*0.001*

*: Questionable items in reliability analysis

Re-test responses were also evaluated in regard to how well the responses corresponsed to the first test. For items 19 and 21, the correct response rates in both tests were below 30%, the incorrect response rate was 50% for item 19 and 38.7% for item 21.

Reliability was further evaluated and analyzed for each item. Table 4 shows percentage of subjects who chose the same option in both tests. In addition, the percentage of "once correct once wrong" responses are shown along with "twice correct" responses. Percentage of "twice wrong" responses (not necessarily by same subjects) were also shown. The overall likelihood to respond correctly or incorrectly varied markedly between items. Some of the items seemed to be easier than others. In items 3, 8, 16 and 22 the "twice-correct" response was obtained almost 80% of the time. In contrast, items 1,7,17,19,21,23 and 33 yielded a twice-wrong response rate of 24-50%. Two of these items, 19 and 21, were not reliable according to the reference sample. The inconsistency of 19 and 21 increased upon test re-test.

The options for each item were classified as true and false responses and evaluated with the Kuder-Richardson 20. For the 34 item test, the KR_{20} was 0.69 and after questions 19 and 21 were excluded, the KR_{20} was 0.72.

DISCUSSION

Our results showed that most of the items in the RMET – Turkish version (34 items) have good reliability. Only two items (19 and 21) were found to be unreliable upon analysis of results. These items showed a low rate of correct responses in both test and re-test. Some items were also found questionable in Hallerback *et al.* (2009)'s study which used items for child version of the RMET. The original RMET child version has 28 items, 25 of which are the same pictures as the adult version. However, the child version includes commonly used words and there is a significant contrast between the target word and foils. One of the items that was found to be unreliable in the Swedish study was the same as item 19 in our study (item 17 in the child version). The other unreliable item (question 21) in our study was not included in the child version and could not be compared. Two other questionable items in the Swedish version (questions 5 and 34 in the Turkish version and 6 and 27 in the Swedish version, respectively) had good reliability in theTurkish version. These findings may be in part explained by the photograph used, methodological difference and perhaps also cultural differences.

As mentioned above, two additional items were excluded from the reliability analysis in accordance to suggestions of the ARC following the pilot studies. The original study was comprised of 40 items, followed by pilot study performed by 8 judges. The items which were responded falsely by at least 4 judges were excluded. After two pilot studies, 36 items were found to be reliable and were included in the test administered to a large number of people (Baron-Cohen et al. 2001). In this study, upon completion of one pilot study, all authors discussed each item and if there was a tendency towards one of the foils, we changed the Turkish counterpart of the target word or foils. This stage was performed meticulously through 3 piloting studies, as any slight difference would lead a change in the responses. In some items we used phrases instead of single words. The target word was "interested" for question 25 but was excluded after the pilot studies. Most of the subjects chose the foil "incredulous" (the foils were incredulous, despondent, panicked). In another excluded item, (item 35) the target word was "nervous". Most of the subjects chose the foil "puzzled" (the foils were puzzled, insisting, comtemplative). In the unreliable question 19, the target word was "tentative" (correctly chosen by 38% of subjects). 26.5% of subjects chose the foil "sarcastic", 24% chose the foil "grateful" and 12% chose the foil "arrogant". In item 21, the target word was "fantasizing" (correctly chosen by 42% of the subjects). 32.5% of the subjects chose the foil "embarrased", 24.8% chose the foil "confused", none chose the foil "panicked". When these four items and foils were evaluated, we observed a tendency towards the foils related to insecurity (incredulous and sarcastic) in items 19 and 25. We did not find a shared theme in items 21 and 35 in relation to other items. Inconsistencies between target words and foils were also detected in 4 items in the original study and these items were excluded before administration. Similarly, 4 items were found to be unreliable in the Swedish study.

Some studies on emotion have investigated the effects of language and culture on interpretation and inferring. Caroll and Russell (1996) reported that British subjects tend to mistake "anger" for "puzzled" when using photos of facial expressions selected from those published by Ekman and Friesen (1976). One approach to the interpretation of emotions is that of *ontological realism* which suggests that emotions are universal natural entities independent of language and culture. In contrast, an approach to emotions as cultural products created through language stems can be described by a philosophical position called *nominalism* (Fernandez-Dolz and Russell 2003). This theoretical controversy in pyschology further necessitates the reliability tests of instruments developed to assess emotions. The cultural influences should be further investigated with the Turkish version of the test.

Consistent with previous studies, the mean scores were significantly higher in female subjects (Baron-Cohen et al. 2001). A gender difference in correct response rates was observed in subjects with primary education while there was no significant difference between female and male subjects with high school or university education (Table 5). This association between ed-

ucation and gender should be further investigated using larger samples. The correct response rates in our study were found to be similar to the rates in the original study (Baron-Cohen et al. 2001) and in the study by Hallerbeck *et al.* (2009). The correct response rates for the general population and college students were 72.77% and 77.77% respectively in the original study (Baron-Cohen et al. 2001). The average overall correct response rate in this study was 73.87% for 117 participants. The study by Hallerbeck et al. included college students and the correct response rate was 78.75%. In our study, the correct response rate was 81.59% for university graduates. These findings illustrate that test performance is comparable to previous studies. Education and gender should still be evaluated as significant variables in studies using the Eyes Test. According to our results, it was found that some items (3 and 18) were correctly answered by almost 90% of the subjects whereas other items (23 and 26) were answered correctly by two thirds of the subjects. Also some foil words (examples are seen in questions 7 and 22) were selected more often while other foil words (examples are seen in 5 and 17) were hardly selected. The Eyes Test has been evaluated solely according to correct answers. This led us consider about a new categorization system according to the valence of foil items. For example, all answers will be assigned a point ranging from 1-4 with the target word corresponding the highest score and others being scored according to the selection frequency of that foil word. In this sense, differences in scores on the Eyes Test can be investigated more precisely and subtle differences can be detected more significantly in clinical or non-clinical study samples. Two previous studies used an additional categorization of RMET items (Harkness *et al.* 2005, Fertuck *et. al.* 2009). In one of these studies items were categorized as negative, positive and neutral after a pilot study in which non-clinical students scored each card using a Likert type scale (1=very negative, 4=neutral, 7=very positive) (Harkness et al. 2005). For future studies this kind of emotional valence scale may be useful especially when investigating emotionality in various patient groups. Response tendencies as shown in emotional valence scale may be important in borderline patients (Fertuck et. al. 2009). Further classification of the items might be obtained by separating and comparing those showing primary emotions from those depicting secondary emotions. As primary emotions are proposed to have distinct neurobiological basis (Pelletier 2003) people with specific psychiatric conditions might perform differently with regards to emotions being primary or secondary. We suggest that a specific group of patients may have a shared tendency towards specific options (foil or target) and this may be associated with different emotional processing of mind reading in distinct clinical or non-clinical groups.

Taken together, this study shows that the Turkish version of the Eyes Test, which contains 34 items, has adequate reliability except for items 19 and 21. After exclusion of these two items, the confidence interval for the difference between the two mean values of the test and retest calculated according to Bland and Altman's method has been found to be adequate for the new 32 items version. The response variance was similar to the original study conducted with non-clinical samples (Baron-Cohen et al. 2001)

In conclusion, the Turkish version of the 32 item Eyes Test, which is widely used in neuroscience studies investigating domains as ToM and emotion regulation, may have utility for clinical or non-clinical samples.

Finally, the major limitation of our study is that the number of participants is small thus may not be representative of the whole Turkish population or a good basis for excluding two items which failed to reach the correct response rate. Further analysis is needed in other pilot studies. Also, only 59.8% of participants underwent re-testing, which may have affected the statistical distribution of results.

REFERENCES

Baron-Cohen S, Jolliffe T, Mortimore C, et al. (1997) Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome. J Child PsycholPsychiatry, 38: 813-822.

Baron-Cohen S, Wheelwright S, Hill J et al. (2001) Reading the Mind in the Eyes Test revised version: a study with normal adults, and adults with Asperger's syndrome or high functioning autism. J Child Psychol Psychiatry, 42: 241-251.

Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1: 307-310.

Bora E, Eryavuz A, Kayahan B et al. (2006) Social functioning, theory of mind and neurocognition in outpatients with schizophrenia; mental state decoding may be a better predictor of social functioning than mental state reasoning. Psychiatry Res, 145: 95 -103.

Bora E, Gokcen S, Kayahan B et al. (2008) Deficits of Social-Cognitive and Social-Perceptual Aspects of Theory of Mind in Remitted Patients With Schizophrenia Effect of residual symptoms. J Nerv Ment Dis, 196: 95-99.

Bora E, Vahip S, Gonul AS et al. (2005) Evidence for theory of mind deficits in euthymic patients with bipolar disorder. Acta Psychiatr Scand, 112: 110-116.

Brüne M, Brüne-Cohrs U (2006) Theory of mind—evolution, ontogeny, brain mechanisms and psychopathology . NeurosciBiobehav Rev, 30: 437-455.

Carroll JM, Russell JA (1996) Do facial expressions signal specific emotions? Judging emotion from the face in context. J Pers. Soc Psychol, 70:205-18.

Corcoran R, Frith CD, Mercer G (1995) Schizophrenia, symptomatology and social inference: Investigating 'theory of mind' in people with schizophrenia. Schizophr Res, 17:5-13.

Cosmides L (1989) The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. Cognition, 31: 187-276.

Craig J, Hatton C, Craig F et al. (2004) Persecutory beliefs, attributions and theory of mind: comparison of patients with paranoid delusions, Asperger's syndrome and healthy controls. Schizophr Res, 69:29-33.

Domes G, Heinrichs M, Michel A et al. (2007) Oxytocin improves "mind-reading" in humans. Biol Psychiatry, 61:731–733.

Drubach DA (2008) The Purpose and Neurobiology of Theory of Mind Functions. J Relig Health, 47:354-365.

Drury VM, Robinson EJ, Birchwood M (1998) Theory of mind' skills during an acute episode of psychosis and following recovery. Psychol. Med, 28:1101-1112.

Dunbar RI (1998) The social brain hypothesis. Evol Anthropol, 6: 178-190.

Erol A, Ünal EK, Gülpek D et al. (2009) Yüzde Dışavuran Duyguların Tanınması ve Ayırt Edilmesi Testlerinin Türk Toplumunda Güvenilirlik ve Geçerlilik Çalışması. Anadolu Psikiyatri Dergisi 10:116-123

Fernández-Dols J, Russell J (2003) Emotion, Affect, and Mood in Social Judgments. Handbook of Psychology, cilt 5, Personality and Social Psychology, T. Millon, MJ Lerner, IB Weiner (Ed). New York. Wiley Press, s.283-298

Fertuck E, Jekal A, Song I et al. (2009) Enhanced 'Reading the Mind in the Eyes' in borderline personality disorder compared to healthy controls. Psychol Med ,39: 1979-88.

Frith CD, Corcoran R (1996) Exploring 'theory of mind' in people with schizophrenia. Psychol Med, 26:521-530.

Hallerbäck MU, Lugnegård T, Hjärthag F et al. (2009) 'The Reading the Mind in the Eyes Test: Test-retest reliability of a Swedish version. Cogn Neuropsychiatry, 14:127-143.

Happé FGE (1994) An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistics, mentally handicapped and normal children and adults. J Autism Dev Disord, 24: 129–154.

Harkness KL, Sabbagh MA, Jacobson JA et al. (2005) Enhanced accuracy of mental state decoding in dysphoric college students. Cognition and Emotion, 19: 999-1025.

Irani F, Platek SM, Panyavin IS et al. (2006) Self-face recognition and theory of mind in patients with schizophrenia and first-degree relatives. Schizophr Res, 88:151-160.

Kelemen O, Erdelyi R, Pataki I et al. (2005). Theory of mind and motion perception in schizophrenia. Neuropsychology, 19: 494-500.

Kelemen O, Keri S, Must A et al. (2004) No evidence for impaired 'theory of mind' in unaffected first-degree relatives of schizophrenia patients. Acta Psychiatr Scand,110:146-149

Langdon R, Coltheart M (1999) Mentalising, schizotypy, and schizophrenia. Cognition, 71: 43-71.

Lee L, Harkness KL, Sabbagh MA et al. (2005) Mental state decoding abilities in clinical depression. J Affect Disord, 86: 247-258

Leslie A (1987) Pretence and representation: The origins of 'theory of mind'. Psychol Rev, 94:412–426.

Murphy D (2006) 'Theory of mind in Asperger's syndrome, schizophrenia and personality disordered forensic patients, Cogn Neuropsychiatry, 11: 99 -111

Pelletier M, Bouthillier A, Levesque J et al. (2003) Separate neural circuits for primary emotions? Brain activity during self-induced sadness and happiness in professional actors. NeuroReport, 14:1111 –1116.

Perner J, Leekham SR, Wimmer H (1987) 3-year-old's difficulty with false belief: The case for a conceptual deficit. Br J Dev Psychol, 5: 125-137.

Premack D, Woodruff G (1978) Does the chimpanzee have a "theory of mind". Behav Brain Sci, 4: 515-526.

Richell R, Mitchell D, Newman C et al. (2003) Theory of Mind and psychopathy: Can psychopathic individuals read the `language of the eyes'. Neuropsychologia, 41:523-526.

Sarfati Y, Hardy-Bayle´ MC, Besche C et al. (1997) Attribution of intentions to others in people with schizophrenia: a non-verbal exploration with comic strip. Schizophr Res, 25: 199–209.

Schmidt KL, Cohn JF (2001) Human facial expressions as adaptations: Evolutionary questions in facial expression research. Am J Phys Anthropol, 33:3-24.

Stone V, Baron-Cohen S, Calder A et al. (2003) Acquired theory of mind impairments in individuals with bilateral amygdala lesions. Neuropsychologia, 41: 209-220.

Torralva T, Kipps CM, Hodges JR et al. (2006) The relationship between affective decision-making and theory of mind in the frontal variant of fronto-temporal dementia. Neuropsychologia, 45: 342-349

Wimmer H, Perner J (1983) Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition, 13: 103-128.