

Published in final edited form as:

Cogn Neuropsychiatry. 2013 July ; 18(4): 326–354. doi:10.1080/13546805.2012.721728.

The “Reading the Mind in the Eyes” test: Systematic review of psychometric properties and a validation study in Italy

Marcello Vellante¹, Simon Baron-Cohen², Mariangela Melis¹, Matteo Marrone¹, Donatella Rita Petretto¹, Carmelo Masala¹, and Antonio Preti^{1,3}

¹Department of Education, Psychology, Philosophy, University of Cagliari, Cagliari, Italy

²Autism Research Centre, Department of Psychiatry, Cambridge University, Cambridge

³Genneruxi Medical Center, Cagliari, Italy

Abstract

Introduction—The “Reading the Mind in the Eyes” test (henceforth, Eyes test) is a simple but advanced Theory of Mind test, and it is widely used across different cultures. This study assessed the reliability and construct (convergent and discriminant) validity of the Eyes test in Italy.

Methods—A sample of 18- to 32-year-old undergraduate students of both sexes ($N = 200$, males = 46%) were invited to fill in the Italian version of the Eyes test, the Empathy Quotient (EQ), the Toronto Alexithymia Scale (TAS), and the Marlowe-Crowne Social Desirability Scale (SDS).

Results—Internal consistency (Cronbach’s alpha) was .605. Confirmatory factor analysis provided evidence for a unidimensional model, with maximal weighted internal consistency reliability = .719. Test–retest reliability for the Eyes test, as measured by intraclass correlation coefficient, was .833 (95% confidence interval = .745 to .902). Females scored significantly higher than males on both the Eyes test and the EQ, replicating earlier work. Those participants who scored lower than 30 on the EQ ($n = 10$) also scored lower on the Eyes test than those who did not ($p < .05$). Eyes test scores were not related to social desirability.

Conclusions—This study confirms the validity of the Eyes test. Both internal consistency and test–retest stability were good for the Italian version of the Eyes test.

Keywords

Autism; Empathy; Reading the Mind in the Eyes; Reliability and validity; Social cognition; Theory of Mind

Introduction

According to the evolutionary psychology framework (Cosmides & Tooby, 1997), emotions are automatic/involuntary response coordination mechanisms, accompanied by distinct types of physiological arousal, selected by evolution to react to basic environmental challenges such as avoiding physical harm (fear), contaminants (disgust), discouraging intentional

damage to basic resources (anger), orienting towards novel cues (surprise), signalling an actual or potential loss (sadness), and attracting companionship or potential mates (joy). In this framework, emotional states are seen as having been shaped by natural selection to enable animals to cope with the threats and opportunities of their physical and social environment (Barrett, Cosmides, & Tooby, 2010; Cosmides & Tooby, 1997).

Emotion recognition is a core component of social cognition, defined as the ability to interpret emotional expressions (in the face, voice, or posture) about others' mental states, to predict their behaviour (Adolphs, 2009; Gallese, Keysers, & Rizzolatti, 2004). In both nonhuman and human primates, the ability to rapidly and correctly detect others' emotional states is crucial for social interaction. In humans, effectiveness in deciphering others' emotions confers an advantage socially and vocationally (Boyatzis & Satyaprasad, 1994; Efenbein, Foo, White, Tan, & Aik, 2007; Leppänen & Hietanen, 2001). Conversely, emotion recognition deficits are implicated in psychopathology (Baron-Cohen, 2003, 2009; Decety & Moriguchi, 2007; Maggini & Raballo, 2004). In particular, the social-cognitive impairment that affects people with autism or schizophrenia involves selective difficulties in emotion recognition (Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Green et al., 2005; Larøi, Fonteneau, Mourad, & Raballo, 2010). Deficits in emotion recognition are reported in other conditions as well, such as anorexia nervosa (Harrison, Tchanturia, & Treasure, 2010), bipolar disorder (Derntl, Seidel, Kryspin-Exner, Hasmann, & Dobmeier, 2009; Ryu, An, Jo, & Cho, 2010), and social anxiety (Machado-de-Sousa et al., 2010). Thus, the investigation of emotion recognition and its contribution to social cognition has clinical relevance, and offers clues to the aetiology of these conditions.

We define emotion recognition as the ability to read subtle cues indicating the emotional state of another person. These cues can be both visual, principally in facial expression, and verbal, principally in prosody, both signalling an internal emotional state. Emotion recognition and the perception of behavioural social cues are integral to social cognition (Adolphs, 2009; Gallese et al., 2004; Gallese, Rochat, Cossu, & Sinigaglia, 2009).

Social cognition encompasses a broader set of abilities involved in the detection and processing of information on goals and intentions of conspecifics. Emotional and behavioural social cues enable us to derive information about others' intentions, beliefs, goals, and desires. The processing of these social cues is sometimes referred to as employing a "theory of mind" (ToM), the ability to attribute mental states to the self and to others, to predict behaviour (Premack & Woodruff, 1978; Samson, 2009).

Different instruments have been developed to assess different components of social cognition. ToM tests generally consist of short stories and questions about them (Bird, Castelli, Malik, Frith, & Husain, 2004; Hallerbäck, Lugnegård, Hjärthag, & Gillberg, 2009; White, Hill, Happé, & Frith, 2009). These tests require intact verbal ability. Instruments testing emotion recognition typically consist of still facial photographs, as in the "Facial emotion identification task" (Ihnen, Penn, Corrigan, & Martin, 1998; Kee, Kern, & Green, 1998), the "Social cognition test" (Hall et al., 2004), or the NimStim Face Stimulus Set (Tottenham et al., 2009). Other tests use vocal stimuli, with the subject invited to identify

emotional states from spoken phrases, as in the “Voice Emotion Identification Test” (Kerr & Neale, 1993) or the “Reading the Mind in the Voice” test (Rutherford, Baron-Cohen, & Wheelwright, 2002). A combination of verbal and dynamic visual stimuli is included in the “Cambridge Mind-reading Face Voice Battery” (Golan, Baron-Cohen, & Hill, 2006). Other tests use videotapes depicting social interactions, and the subject is asked about the nature and the quality of these interactions. Examples are the “Social Cue Recognition Test” (Ihnen et al., 1998) and the “Videotape Emotion Identification Test” (Kee et al., 1998). A finer but more time-consuming evaluation of emotion recognition is based on the assessment of computerised distorted facial pictures (morphing) (Scrimin, Moscardino, Capello, Altoè, & Axia, 2009). In all these tests the participant is asked to identify an emotion or to rate its intensity across a range of expressions.

The simplest emotion recognition test and one of the most widely used is the “Reading the Mind in the Eyes” test (henceforth, the Eyes test), which underwent a revision in the adult version, and which includes 36 still pictures of the eye regions illustrating emotionally charged or neutral mental states (Baron-Cohen et al., 1997; Baron-Cohen, Wheelwright, Hill, et al., 2001). The core of the Eyes test involves the matching of the semantic definition of a mental state (e.g., “worried”, “annoyed”) to the picture of the eye-region expression displayed in the screen. This is assumed to involve an unconscious, automatic and rapid matching of past memories concerning similar expressions with a lexicon of mental state terms (Baron-Cohen, Wheelwright, Hill, et al., 2001).

The Eyes test is considered an advanced ToM test, since participants have to put themselves into the mind of the person shown in the photograph, and attribute a relevant mental state to them. However, because the judgements can only be on the basis of facial expression, the test can also be considered an emotion recognition test. The Eyes test has been used in samples of people expected to show impairment of social cognition, such as people with autism, schizophrenia, and anorexia (Baron-Cohen, Wheelwright, Hill, et al., 2001; Harrison et al., 2010; Kettle, O’Brien-Simpson, & Allen, 2008). The Eyes test has also been used to assess the impact of pharmacological treatment on social cognition, in particular to assess the benefit of oxytocin administration in both typical individuals (Domes, Heinrichs, Michel, Berger, & Herpertz, 2007), and in people with autism (Guastella et al., 2010) or schizophrenia (Feifel et al., 2010). On the Eyes test, deficits are found in people with autism (Baron-Cohen, Wheelwright, Hill, et al., 2001), schizophrenia (Kettle et al., 2008; Köther et al., 2012; Schimansky, Rössler, & Haker, 2012), eating disorders, particularly those with anorexia nervosa (Harrison et al., 2010), and in depressed elderly people with a history of suicide attempt (Szanto et al., 2012). Patients with Parkinson’s disease (Tsuruya, Kobayakawa, & Kawamura, 2011) and Huntington’s disease (Allain et al., 2011) also show impairment on the Eyes test. Patients with borderline personality disorder may show enhanced performance compared to typical controls (Fertuck et al., 2009; Schilling et al., 2012), perhaps reflecting heightened attention to affective threat cues in the eye stimuli.

The ability to decode the feelings and thoughts of others from the eyes develops at a young age (Baron-Cohen, Wheelwright, Spong, Scahill, & Lawson, 2001; Gunther Moor et al., 2012). Neuroimaging studies suggest that the Eyes test causes activation of a network including the posterior superior temporal sulcus (pSTS) and the inferior frontal gyrus

(Adams et al., 2009; Baron-Cohen et al., 1999, 2006; Gunther Moor et al., 2012). However, various higher order ToM tasks may be related to differing information-processing components, since different ToM tests may tap different domains. For example, Ahmed and Miller (2011) found lack of correlation between the Eyes test, a revised version of the Strange Stories test (Happé, 1994), and the Faux pas test (Gregory et al., 2002) in a sample including 123 typical people. Other studies have found the Eyes test to be highly correlated with the Strange Stories test (Baron-Cohen, Wheelwright, Hill, et al., 2001).

The Eyes test is based on facial pictures of Caucasian people, and has been used in predominantly Anglo-Saxon countries (Goding, Johnson, & Peterman, 2010; Guastella et al., 2010; Harrison et al., 2010; Irani et al., 2006; Kettle et al., 2008; Valla et al., 2010). There is evidence that facial pictures of American, Caucasian people expressing the six basic emotions can be readily identified and correctly labelled across cultures (Ekman, Sorenson, & Friesen, 1969; Izard, 1971). However, cultural differences have been reported in emotion recognition of complex stimuli based on facial pictures of Caucasian American people (Izard, 1971), and cultural differences have been reported even when using stimuli that are culturally compatible with the participant being tested (Elfenbein et al., 2002). Cultural differences may be even larger for complex mental states as those depicted in the Eyes test, but detailed studies on this topic are lacking. Studies based on the Turkish, Hungarian, Japanese, French, German, and Spanish versions of the Eyes test have been published so far (Bora et al., 2005; de Achávala et al., 2010; Havet-Thomassin, Allain, Etcharry-Bouyx, & Le Gall, 2006; Kelemen, Keri, Must, Benedek, & Janka, 2004; Kunihiro, Senju, Dairoku, Wakabayashi, & Hasegawa, 2006; Voracek & Dressler, 2006). However, most of these studies did not detail information on the psychometric properties of the test in these cultures.

Information on reliability of the Eyes test are rarely reported. Two studies found low internal coherence, as measured by Cronbach's alpha (Harkness, Jacobson, Duong, & Sabbagh, 2010; Voracek & Dressler, 2006), and two other studies found acceptable internal coherence as assessed with Cronbach's alpha (Dehning et al., 2012) or Guttman's split-half method (Serafin & Surian, 2004). Test-retest stability was found acceptable for both the adult version (Yildirim et al., 2011) and the child version (Hallerbäck et al., 2009) of the Eyes test. However, independent replication of these findings is lacking (Table 1).

The Eyes test is usually used as a total score resulting from the sum of the correct guess on the 36 items, but the monodimensional structure of the test has never been formally investigated by factor analysis. The most replicated finding in past studies is the greater ability of female participants on the Eyes test. In six out of 17 studies, a statistically significant female advantage was observed, with Cohen's *d* ranging from 0.22 to 0.94. Unfortunately, scores by gender are often lacking and only the statistical results are reported, which makes a meta-analysis of findings complicated.

Eyes test scores are often reported to correlate with self-reported empathy, but two studies failed to find a link with self-report measures of empathy. No link with the Interpersonal Reactivity Index (Davis, 1983) was found in the Muller et al. (2010) study, and only the social skills factor of the EQ was modestly related to the Eyes test in the Lawrence et al.

(2004) study. Performance on the Eyes test requires intact verbal abilities, and a link with IQ, and verbal IQ in particular, was found in four out of eight studies (see Table 1).

One study examined culture differences on the Eyes test in a mixed American and Japanese sample, using both white American and Asian eyes stimuli. Authors found that both Japanese and white American participants performed better on same-culture (American participants vs. American eyes stimuli; Asian participants vs. Asian eyes stimuli) compared to other-culture stimuli, such as American participants versus Asian eyes stimuli or the reverse (Adams et al., 2009). Because of the problems with the identification of the Caucasian eyes stimuli, the reaction time and not the number of correct replies is considered in the Japanese version of the Eyes test (Kunihira et al., 2006). In a study carried out in Ethiopia, 237 medical students showed low scores on the Caucasian-based original Eyes test: on average below 22 as against 25–28 mean scores in typical Western student samples (see Table 1). Another study by Baron-Cohen et al. (1996) examined complex mental states as depicted in drawings of faces of a contemporary British artist (Hockney) and a seventeenth-century Spanish painter (Velazquez). In the Baron-Cohen et al. study, people from three different cultures (Britain, Spain, and Japan) showed good agreement in their judgement of most depicted mental states from facial information. However, cultural differences were found on three out of the 11 mental states: scheme, wariness, and guilt. Moreover, although there is evidence that members of different cultures reliably judge spontaneously expressed emotions, culturally influenced expressions are often displayed after initial, immediate, and universal emotional displays (Matsumoto, Olide, Schug, Willingham, & Callan, 2009; Matsumoto, Willingham, & Olide, 2009).

In Latin cultures, greater emphasis is placed on the expression of emotions than in Anglo-Saxon countries (Bhugra & McKenzie, 2003). This greater emphasis on emotional display might be related to differences in the ability of recognising emotions and complex affective states in others. The present study set out to assess reliability and construct (convergent and discriminant) validity of the Eyes test in Italy, a Southern European Mediterranean culture, to further extend its psychometric properties. We expected the Italian version of the Eyes test to display acceptable reliability, as far as internal coherence and test–retest stability are concerned, to show positive associations with measures of empathy and negative associations with measures of alexithymia, and to be unrelated to social desirability, differently from self-report measures of empathy.

Methods

Participants

Participants were 200 undergraduate students attending the courses of the University of Cagliari (Italy). Participants were enrolled from a wide array of courses: engineering ($n = 45$), psychology ($n = 57$), mathematics ($n = 2$), architecture ($n = 6$), political science ($n = 15$), medicine ($n = 4$), languages ($n = 9$), education ($n = 13$), classical studies ($n = 19$), pharmacy ($n = 4$), biology ($n = 5$), economics ($n = 12$), law ($n = 7$), geology ($n = 1$), physics ($n = 1$). This sample's participation rate was 91.3%, from 219 subjects invited to take part in the study (seven refused to take part in the study; six refused to take the laboratory test after having completed the questionnaires; six did not complete the questionnaires). Participation

was voluntary and no fee or other incentive was provided for taking part in the study. The institutional review board of the Department of Psychology of the University of Cagliari approved this study. The protocol of the research project conformed to the guidelines of the 1995 Declaration of Helsinki (as revised in Tokyo 2004). All participants provided informed consent.

Measures

Each participant was told the data would remain confidential, and received a booklet containing the questionnaires listed next, which they were asked to complete.

The Empathy Quotient (EQ)—The EQ is a 60-item questionnaire, with 40 questions tapping empathy and 20 filler items. Responses are given on a 4-point Likert scale. Scores can range from 0 to 80 (Baron-Cohen & Wheelwright, 2004; Lawrence et al., 2004). A cutoff score of fewer than 30 is useful to differentiate adults with autism from controls. We used the Italian version of the questionnaire (Preti et al., 2011), which shows acceptable internal consistency, concurrent and convergent validity, and good test–retest reliability.

The 20-item Toronto Alexithymia Scale (TAS-20)—The TAS-20 is a self-report scale that is comprised of 20 items, and it is the most frequently used self-assessment instrument to assess alexithymia because of its good reliability and construct validity (Bagby, Taylor, & Parker, 1994; Taylor, Bagby, & Parker, 2003; Taylor et al., 1988). We used the validated Italian version of the questionnaire (Bressi et al., 1996). Since the factorial structure of the TAS-20 has not been always supported (Müller, Buhner, & Ellgring, 2003), we used the total score only to assess alexithymia. Some studies report a high degree of alexithymia in individuals with autism (Berthoz & Hill, 2005; Bird et al., 2010; Lombardo et al., 2007).

The Marlowe-Crowne Social Desirability Scale (SDS)—This is a 33-item self-report questionnaire aimed at measuring social desirability (Crowne & Marlowe, 1960). Subjects rate the extent to which they agree (True) or disagree (False) with each item: the 15 items keyed False are likely but socially undesirable and are thought to measure denial and self-deception (e.g., “I am sometimes irritated by people who ask me favours”); the 18 items keyed True are improbable but socially desirable and are thought to measure a tendency to positive attribution, or to attribute to the self traits that are seen by society positively (e.g., “No matter who I’m talking to, I am always a good listener”) (Ramaiah, Schill, & Leung, 1977). The Italian version of the SDS showed good psychometric functioning when tested in nonclinical populations and, as in past studies (Lane et al., 1990), showed a negative correlation with measures of psychopathology (Miotto, De Coppi, Frezza, Rossi, & Preti, 2002; Miotto & Preti, 2008).

General sociodemographic information from self-report data was collected for the following variables: age, sex, and socioeconomic status. As a measure of socioeconomic status we used the highest level of parental education (Galobardes et al., 2006), which was further classified into three categories: lower than high school diploma, high school diploma, college graduate or higher.

After completion of the booklet containing the questionnaires, each participant was invited to sit in a quiet room where s/he took the Reading the Mind in the Eyes test (Eyes test; Baron-Cohen, Wheelwright, Hill, et al., 2001). The adult, 36-item revised version of the Eyes Test was administered in the Italian version. We used the version of the Eyes test as reported in the Italian translation of *The Essential Difference* (Baron-Cohen, 2003, 2004); this version was based on translation and backtranslation of the descriptors, with a final check by the author.

Participants were randomly presented with a series of 36 photographs of the eye region of 19 actors and 17 actresses. Each photo was surrounded by four single-word mental state descriptors, e.g., bored, angry, one at each corner. One of these descriptors targeted the mental state depicted in the photo, and the others were foils. The Eyes test is based on a four-alternative forced-choice paradigm, with 25% correct guess rate. On each item, a response rate equal or above 50% indicates that participants are really detecting mental states and are not making random guess. Participants were instructed to choose which of the four descriptors best describes what the person in the photo is thinking or feeling. Participants could take as long as they wanted, and progressed to the next item when ready. They also could request an explanation of the meaning of the descriptors and could consult a prepared glossary if they were unsure of the definition of any descriptor word used. Score on the test is the number of descriptors correctly identified by the participants, i.e., the number of mental states correctly identified. The maximum score is 36.

Statistical analysis

All data were coded and analysed using the Statistical Package for Social Sciences (SPSS) for Windows (Chicago, IL 60606, USA), version 13. All tests were two-tailed, with $\alpha = .05$. Categorical data were analysed in intergroup comparisons with χ^2 , or Fisher's exact test, when appropriate ($n < 5$ in any cell in binary comparison). Student *t*-test or the Mann-Whitney test, when appropriate, was used to compare the ordinal variables. Pearson's *r* correlation coefficients were used to examine associations between two variables.

Scale reliability was measured by Cronbach's coefficient alpha, a measure of internal consistency. For the Eyes test, reliability was also measured by maximal weighted internal consistency reliability from confirmatory factorial analysis (CFA) testing the unidimensional model (see later). To compare groups, reliability values of .70 are considered satisfactory (Bland & Altman, 1997).

Test-retest stability of the Eyes test was evaluated in a subgroup of 36 participants (one per item), who were invited to fill in the Eyes test again after 30 days from initial assessment. Follow-up completion rate was 80% for the test-retest reliability sample (36 out of 45 contacted: five refused to take part in the study because they did not have time; three accepted to take part in the study but did not show up on the day of the test; one took the test but did not return for the retest). Test-retest stability was assessed using intraclass correlation coefficient (ICC), with 95% confidence interval (CI). The ICC is dimensionless statistics that describes the reproducibility of repeated measurements in the same population: ICC values .60 are considered acceptable for clinical use (Brennan & Silman, 1992). The Bland and Altman (1986) method was used, too, to assess agreement at retest. The Bland-

Altman plot visualises the agreement between the scores of a test measured at two different assessment points by plotting the difference between test and retest scores against the mean value of test and retest scores for each participant. Confidence intervals for the mean difference are calculated to determine if the mean difference deviates significantly from zero, which should not occur. Graphically, the upper and lower limits of agreement are drawn, indicating the range within which 95% of the test scores of two assessments can be expected to vary.

The Eyes test is supposed to discriminate people with ToM deficits from typical people, and for this a summary score is often used. Therefore, the unidimensional model, assuming all items loaded on a single factor, was tested by CFA, which was carried out with EQS for Windows (Bentler, 2008), version 6.1. An alternative three-factor model, based on an items valence classification, was also tested, by distinguishing items with positive (eight items; i.e., playful, fantasising [Items 6 and 21], thoughtful, friendly, interested, flirtatious, confident), negative (12 items, i.e., upset, worried, regretful, accusing, doubtful, preoccupied [Item 9], defiant, hostile, cautious, distrustful, nervous, suspicious), and neutral (16 items, i.e., desire, insisting, uneasy, despondent, preoccupied [Item 22], cautious, sceptical, anticipating, contemplative, decisive, tentative, pensive, interested, reflective, serious, concerned) valence (Harkness, Sabbagh, Jacobson, Chowdrey, & Chen, 2005). Maximum likelihood under elliptical distribution was used to estimate fit. Chi-square is the traditional fit index for evaluating overall model fit, since it assesses the magnitude of discrepancy between the sample and fitted covariances matrices (Hu & Bentler, 1999, p. 2). However, the use of the chi-square likelihood ratio test to assess model fit is unsatisfactory for a number of reasons (Tanaka, 1993), including sensitivity to sample size. The ratio chi-square/degrees of freedom (*df*) was calculated in addition to the chi-square to evaluate model fitting, with ratios larger than 2 indicating poor fit (Byrne, 1989). In this study, we also used other parameters for fit estimate, namely: the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardised root mean square residual (SRMR). RMSEA = .06, SRMR = .09, and CFI = .90 are considered acceptable (Browne & Cudeck, 1993; Hu & Bentler, 1999).

Convergent validity of the Eyes test was evaluated by testing the association of Eyes test scores with scores on the EQ and the TAS. The Eyes test was expected to covary with self-reported measures of empathy, such as the EQ, and to be negatively related to measures of emotion recognition deficits, such as the TAS. Discriminant validity of the Eyes test was evaluated by examining the links with the SDS. The Eyes test was expected to be unrelated to social desirability, as measured by the SDS.

Sample size

We carried out a preliminary power calculation to determine the minimum sample necessary to test our hypothesis. Based on a Pearson product-moment correlation test, a minimum sample of 82 participants would be needed to achieve 80% power to detect $r = .30$ (medium effect size) in the relationship between scores on the Eyes test and those of the measures of empathy or alexithymia at a two-sided significance level of .05. Using a nonparametric Spearman correlation index a sample of 104 participants would be needed, at the same levels

of power and significance, since efficiency of the Spearman coefficient is about .91 of the corresponding Pearson coefficient (Lehmann, 1975). Calculation was performed using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007).

Results

The final sample included 200 participants (males = 46%). The mean age in the sample was 24.1 years \pm 2.8 (range 19 to 32 years old), with no sex differences in age (males: 24.3 \pm 2.7; females: 23.8 \pm 2.9; $t = 1.22$, $df = 198$, $p = .223$). There was no sex difference in having a parent with a university degree (males $n = 17$; 18.5%; females, $n = 12$; 11.1%). Among participants, none was married (Table 2).

Reliability analysis

Internal consistency, as measured by Cronbach's alpha, was .800 for the EQ, .743 for the TAS, .723 for the SDS, and .605 for the Eyes test.

Factorial analysis of the Eyes test

The unidimensional model proved acceptable on the basis of fit indexes, but for the CFI: Chi-square = 675.69, $df = 594$, $p = .011$, Chi-square/degrees of freedom = 1.1, CFI = .810, RMSEA = .026 (95% CI = .014 to .036), SRMR = .068. Maximal weighted internal consistency reliability for the unidimensional model was .719. The three-factor model did not converge with any adjustment in the distribution. A forced three-factor model by unweighted least square based on polychoric correlations was carried out with the FACTOR program (version 8.02; Lorenzo-Seva & Ferrando, 2006). The fit of the solution was poor: Chi-square=2017.98, $df = 525$, $p < .0001$, Chi-square/degrees of freedom = 3.8, CFI = .310, SRMR = .0716, RMSEA not provided. Item loading on the three factors did not follow the expected distribution. Items expected to belong to the "positive valence" or the "negative valence" factor loaded indifferently on each of the extracted three factors.

Distribution of response on the Eyes test

Females scored significantly higher than males on the Eyes test ($t = -2.50$, $df = 198$, $p = .013$, Hedges' $g = 0.34$, 95% CI = 0.06 to 0.62); no other differences on the sociodemographic variables were found (Table 2). Since the scores were not normally distributed (Shapiro-Wilk test test: $p < .001$), differences on the Eyes test by gender were reanalysed with the nonparametric Mann-Whitney U-Test, which confirmed the difference, $z = -2.09$, $p = .036$. Overall, females were more able than males to correctly identify the target descriptor over the foils (Table 3). In two cases only the target descriptor was selected by less than 50% of the sample (Items 7 and 35). In Item 7, about 25% participants selected the foil word "friendly" rather than the correct word "uneasy". In Item 35, participants were as likely to select the foil word "puzzled" (34%) as the correct word "nervous" (36.5%). These items can be considered difficult items, with lower margin for differentiation.

Figure 1 summarises the distribution of the Eyes test scores in the sample. As can be seen, a small fraction of the sample was particularly poor in the identification of the target descriptor: 15 subjects (7.5%) scored less than 20 on the Eyes test, corresponding to 1

standard deviation below the mean. This group can be said to have “low empathy” on the Eyes test. Males were more likely to have low empathy on the Eyes test: 13 (14.1%) versus 2 (1.9%), Fisher’s exact test $p = .002$; Cramér’s $V = .232$.

Measures of empathy and alexithymia

Females scored higher than males on the EQ ($t = -5.34$, $df = 198$, $p < .0001$, Hedges’ $g = 0.76$, 95% CI = 0.47 to 1.05). There were no differences by gender on the TAS or on the SDS (Table 4). In the sample 1 female (0.9%) but 9 male (9.8%) participants scored lower than 30 on the EQ, the cut-off score that best differentiates Autism Spectrum Conditions from controls (Fisher’s exact test $p = .006$). In the sample 3 female (2.8%) and 6 male (6.5%) participants scored higher than 61 on the TAS, the threshold that suggests the presence of alexithymia (Fisher’s exact test $p = .30$).

Eyes test, EQ, TAS, and SDS

In females, scores on the Eyes test did not correlate with scores on the EQ, the TAS or the two subscale of the SDS (Table 4, bottom). In males, scores on the Eyes test were negatively related to TAS scores, but not to EQ or SDS scores. EQ was negatively related to the TAS and positively related to the SDS, as expected; the TAS was negatively related to the SDS, again as expected. Those participants who scored lower than 30 on the EQ ($n = 10$) also scored lower on the Eyes test than the remaining participants ($n = 190$): 21.9 ± 5.1 vs. 25.0 ± 4.1 , $t = 2.32$, $df = 198$, $p = .021$ (Hedges’ $g = -0.74$, 95% CI = -1.38 to 0.10). People with “alexithymia” scores on the TAS ($n = 9$) scored lower on the Eyes test than those in the nonalexithymia interval ($n = 191$), but the difference was not statistically significant: 22.6 ± 3.3 vs. 24.9 ± 4.2 , $t = 1.62$, $df = 198$, $p = .107$ (Hedges’ $g = -0.55$, 95% CI = -1.22 to 0.12).

Test–retest stability at 1-month interval

In the subgroup involved in the test–retest assessment ($n = 36$), the mean score on the Eyes test was 23.8 ($SD = 5.2$) at the test and 25.1 (4.7) at the retest. Test–retest reliability for the Eyes test, as measured by ICC, was .833 (95% CI = .745 to .902). The mean difference between the first and the second assessment with the Eyes test in the 36 participants was 1.3 ($SD = 4.4$). The 95% CI for the mean difference was -0.18 to 2.79 (i.e., 0 is within the confidence interval, therefore the mean difference did not differ statistically from 0). By plotting the differences and the means of the two assessments in the Bland-Altman plot, two cases out of 36 (5.6%) were outside the upper and lower limits of agreement (Figure 2).

Discussion

This study provides new evidence confirming the validity of the “Reading the Mind in the Eyes” test: Typical participants who score low on the Empathy Quotient (EQ) also scored lower on the Eyes test. Female participants scored higher than male participants on the EQ and the Eyes test, further providing evidence of validity of the Eyes test. This confirms the earlier reports of females scoring higher than males on tests of empathy (Baron-Cohen & Wheelwright, 2004; Lawrence et al., 2004; Von Horn, Bäckman, Davidsson, & Hansen, 2010; but see Eisenberg & Lennon, 1983). More importantly, Eyes test scores were not

related to social desirability: that is, the Eyes test is free from the social desirability bias, as expected by its quasiecological nature.

As far as reliability is concerned, there is some agreement that Cronbach's coefficient alpha is a poor index of unidimensionality, and that how much variance in a test is accounted for by one common factor might be a better index of the internal coherence of a test (Bentler, 2009; Sijtsma, 2009; Zinbarg, Yovel, Revelle, & McDonald, 2006). Reliability of the Eyes test was rarely reported in past studies. Voracek and Dressler (2006) found that Cronbach's alpha was .63 in men and .60 in women. Harkness et al. (2010) found Cronbach's alpha = .58 in a sample including 93 college students. Dehning et al. (2012) found Cronbach's alpha = .70 in 237 Ethiopian medical students, despite a low mean score on the test (on average, below 22). By using a different Italian version of the adult Eyes test than the one used in this study, Serafin and Surian (2004) found a reliability value of .77, as measured by the Guttman split-half technique. In the present study, Cronbach's alpha was .60, whereas maximal weighted internal consistency reliability for the unidimensional model provided a better estimate (.719), indicating acceptable internal consistency of the Eyes test in Italy.

Test-retest stability was reported to be acceptable in a past study from Turkey, with ICC = .65 (Yildirim et al., 2011). By using the child Swedish version of the Eyes test (28 items), Hallerbäck et al. (2009) found evidence of stability of the tool, according to the Bland and Altman (1986) method. We found good test-retest stability of the Italian version of the Eyes test, with ICC = .833, and good evidence of repeatability of the measurement according to the Bland and Altman method.

To our knowledge, this is the first study providing evidence that the Eyes test can be effectively scored as a single factor. Factor loading was far from being optimal: 21 items out of 36 did not reach the minimal acceptable threshold for factor loading (.25, i.e., around 5% of variance explained). Past studies used composite scores of the Eyes test, by subdividing items on the basis of their content. Three subscores were used in some studies, for "positive", "negative", and "neutral" items. We tested this three-factor structure of the Eyes test but the model was rejected on the basis of fit indexes and factor loading. This may be because the mental states in the test are sufficiently different to each other, so that they cannot easily be grouped into a small number of categories.

Past studies report consistent evidence for a correlation between EQ and Eyes test scores. One study also found a negative correlation ($r = -.43$) between TAS and Eyes test scores in a sample including 30 individuals with high-functioning autism or Asperger syndrome and 30 sex- and age-matched typical controls (Lombardo, Barnes, Wheelwright, & Baron-Cohen, 2007). In this study we did not find a correlation between EQ and Eyes test scores in both males and females. Moreover, a link with TAS scores was found in males only, and this may reflect that empathy functions differently between the two sexes (Valla et al., 2010). Males are more likely to display higher levels of autistic traits (Baron-Cohen, Wheelwright, Skinner, Martin, & Clublely, 2001), and a higher chance of alexithymia is reported in people with autism (Berthoz & Hill, 2005; Bird et al., 2010; Lombardo et al., 2007). Those who scored lower than 30 on the EQ, the cutoff score that best differentiates autism from

controls, also had statistically lower scores on the Eyes test, providing some evidence for an association between Eyes test performance and autistic traits.

Analysis of single items also provided evidence for their discriminative properties: Some of the emotions depicted in the pictures were easily detected by the majority of participants, whereas other items were more difficult. Two items only were identified by less than 50% of participants. It has been proposed that the degree of difficulty of an item might vary from the original to the translated version because of cultural differences in the meaning attributed to the target definitions (Hallerbäck et al., 2009). This hypothesis can be tested by a comparative study only.

One limitation of the study is the inclusion of undergraduate students only. The lack of more varied strata of the population precludes the establishment of baseline for the general population. In particular, no evaluation of education level impact on scores can be calculated using undergraduate students only. Although previous studies have not found major effects of education or IQ on this test (Baron-Cohen, Wheelwright, Hill, et al., 2001), some studies have found an association between IQ and the Eyes test (Ahmed & Miller, 2011; Golan & Baron-Cohen, 2006; Kenyon et al., 2012; Stanford et al., 2011). However, in a sample of undergraduate students it is unlikely that IQ would show variations large enough to be detected statistically.

In conclusion, the “Reading the Mind in the Eyes” test is a freely available measure of facial affect recognition, easy to administer and easy to score. It has been translated into many languages, and this allows comparability. This study shows that the Eyes test, in its revised, adult version, is a valid and reliable measure, which can be used to assess differences in recognition of complex affective states, including nonbasic emotions.

Acknowledgments

SBC was supported by the MRC UK and the Wellcome Trust during the period of this study.

References

- Adams RB Jr, Rule NO, Franklin RG Jr, Wang E, Stevenson MT, Yoshikawa S, et al. Ambady N. Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience*. 2009; 22:98–108.
- Adolphs R. The social brain: Neural basis of social knowledge. *Annual Reviews of Psychology*. 2009; 60:693–716.
- Ahmed FS, Miller SL. Executive function mechanisms of theory of mind. *Journal of Autism and Developmental Disorders*. 2011; 41:667–678. [PubMed: 20811770]
- Alaerts K, Nackaerts E, Meyns P, Swinnen SP, Wenderoth N. Action and emotion recognition from point light displays: An investigation of gender differences. *PLoS One*. 2011; 6
- Ali F, Chamorro-Premuzic T. Investigating Theory of Mind deficits in nonclinical psychopathy and Machiavellianism. *Personality and Individual Differences*. 2010; 49:169–174.
- Allain P, Havet-Thomassin V, Verny C, Gohier B, Lancelot C, Besnard J, et al. Le Gall D. Evidence for deficits on different components of theory of mind in Huntington’s disease. *Neuropsychology*. 2011; 25:741–751. [PubMed: 21728429]
- Bagby RM, Taylor GJ, Parker JD. The twenty-item Toronto Alexithymia Scale–II. Convergent, discriminant, and concurrent validity. *Journal of Psychosomatic Research*. 1994; 38:33–40.

- Baron-Cohen, S. *The essential difference: Men, women and the extreme male brain*. London, UK: Penguin; 2003.
- Baron-Cohen, S. *Questione di cervello: La differenza essenziale tra donne e uomini*. Milano, Italy: Mondadori; 2004.
- Baron-Cohen S. Autism: The empathizing-systemizing (E-S) theory. *Annals of the New York Academy of Sciences*. 2009; 1156:68–80. [PubMed: 19338503]
- Baron-Cohen S, Wheelwright S. The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*. 2004; 34:163–175. [PubMed: 15162935]
- 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. *Journal of Child Psychology and Psychiatry*. 1997; 38:813–822. [PubMed: 9363580]
- Baron-Cohen S, Ring H, Chitnis X, Wheelwright S, Gregory L, Williams S, et al. Bullmore E. fMRI of parents of children with Asperger syndrome: A pilot study. *Brain and Cognition*. 2006; 61:122–130. [PubMed: 16460858]
- Baron-Cohen S, Ring HA, Wheelwright S, Bullmore ET, Brammer MJ, Simmons A, Williams SC. Social intelligence in the normal and autistic brain: An fMRI study. *European Journal of Neuroscience*. 1999; 11:1891–1898. [PubMed: 10336657]
- Baron-Cohen S, Riviere A, Cross P, Fukushima M, Bryant C, Sotillo M, et al. French D. Reading the Mind in the Face: A cross-cultural and developmental study. *Visual Cognition*. 1996; 3:39–59.
- 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. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. 2001; 42:241–251.
- Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/highfunctioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*. 2001; 31:5–17. [PubMed: 11439754]
- Baron-Cohen S, Wheelwright S, Spong A, Scahill V, Lawson J. Are intuitive physics and intuitive psychology independent? *Journal of Developmental and Learning Disorders*. 2001; 5:47–78.
- Barrett HC, Cosmides L, Tooby J. Coevolution of cooperation, causal cognition and mindreading. *Communicative and Integrative Biology*. 2010; 3:522–524. [PubMed: 21331228]
- Bentler, PM. EQS 6.1 structural equations program manual. Encino, CA: Multivariate Software; 2008.
- Bentler PM. Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*. 2009; 74:137–143. [PubMed: 20161430]
- Berthoz S, Hill EL. Reliability of the Bermond-Vorst Alexithymia Questionnaire: Data from adults with autism spectrum disorder, their relatives and normal controls. *European Psychiatry*. 2005; 20:291–298. [PubMed: 15935431]
- Bhugra D, McKenzie K. Expressed emotion across cultures. *Advances in Psychiatric Treatment*. 2003; 9:342–348.
- Bird CM, Castelli F, Malik O, Frith U, Husain M. The impact of extensive medial frontal lobe damage on “Theory of Mind” and cognition. *Brain*. 2004; 127:914–928. [PubMed: 14998913]
- Bird G, Silani G, Brindley R, White S, Frith U, Singer T. Empathic brain responses in insula are modulated by levels of alexithymia but not autism. *Brain*. 2010; 133:1515–1525. [PubMed: 20371509]
- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 1:307–310. [PubMed: 2868172]
- Bland JM, Altman DG. Statistics notes: Cronbach’s alpha. *British Medical Journal*. 1997; 314:572. [PubMed: 9055718]
- Bora E, Vahip S, Gonul AS, Akdeniz F, Alkan M, Ogut M, Eryavuz A. Evidence for theory of mind deficits in euthymic patients with bipolar disorder. *Acta Psychiatrica Scandinavica*. 2005; 112:110–116. [PubMed: 15992392]
- Boyatzis CJ, Satyaprasad C. Children’s facial and gestural decoding and encoding: Relations between skills and with popularity. *Journal of Nonverbal Behavior*. 1994; 18:37–55.

- Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. *British Medical Journal*. 1992; 304:1491–1494. [PubMed: 1611375]
- Bressi C, Taylor G, Parker J, Bressi S, Brambilla V, Aguglia E, et al. Invernizzi G. Cross validation of the factor structure of the 20-item Toronto Alexithymia Scale: An Italian multicenter study. *Journal of Psychosomatic Research*. 1996; 41:551–559. [PubMed: 9032718]
- Browne, MW, Cudeck, R. Alternative ways of assessing model fit Testing structural equation models. Bollen, KA, Long, JS, editors Newbury Park, CA: Sage; 1993. 136–161.
- Byrne, BM. A primer of LISREL: Basic applications and programming for confirmatory factor analytic model. New York, NY: Springer-Verlag; 1989.
- Carroll JM, Chiew KY. Sex and discipline differences in empathising, systemising and autistic symptomatology: Evidence from a student population. *Journal of Autism and Developmental Disorders*. 2006; 36:949–957. [PubMed: 16897399]
- Chapman E, Baron-Cohen S, Auyeung B, Knickmeyer R, Taylor K, Hackett G. Fetal testosterone and empathy: evidence from the empathy quotient (EQ) and the “reading the mind in the eyes” test. *Social Neuroscience*. 2006; 1:135–148. [PubMed: 18633782]
- Cook CM, Saucier DM. Mental rotation, targeting ability and Baron-Cohen’s Empathizing–Systemizing Theory of Sex Differences. *Personality and Individual Differences*. 2010; 49:712–716.
- Cosmides L, Tooby J. Dissecting the computational architecture of social inference mechanisms. *Ciba Foundation Symposium*. 1997; 208:132–156. [PubMed: 9386910]
- Crowne D, Marlowe D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*. 1960; 24:1397–1403.
- Davis MH. Measuring individual differences in empathy: Evidence for a multi-dimensional approach. *Journal of Personality and Social Psychology*. 1983; 44:113–126.
- de Achávala D, Costanzo EY, Villarreal M, Jáuregui IO, Chiodi A, Castro MN, et al. Guinjoan SM. Emotion processing and theory of mind in schizophrenia patients and their unaffected first-degree relatives. *Neuropsychologia*. 2010; 48:1209–1215. [PubMed: 20026084]
- Decety J, Moriguchi Y. The empathic brain and its dysfunction in psychiatric populations: Implications for intervention across different clinical conditions. *BioPsychoSocial Medicine*. 2007; 1
- Dehning S, Girma E, Gasperi S, Meyer S, Tesfaye M, Siebeck M. Comparative cross-sectional study of empathy among first year and final year medical students in Jimma University, Ethiopia: Steady state of the heart and opening of the eyes. *BMC Medical Education*. 2012; 12:34. [PubMed: 22624580]
- Derntl B, Seidel EM, Kryspin-Exner I, Hasmann A, Dobmeier M. Facial emotion recognition in patients with bipolar I and bipolar II disorder. *British Journal of Clinical Psychology*. 2009; 48:363–375. [PubMed: 19220936]
- Domes G, Heinrichs M, Michel A, Berger C, Herpertz SC. Oxytocin improves “mind-reading” in humans. *Biological Psychiatry*. 2007; 61:731–733. [PubMed: 17137561]
- Eisenberg N, Lennon R. Sex differences in empathy and related capacities. *Psychological Bulletin*. 1983; 94:100–131.
- Elfenbein HA, Mandal MK, Ambady N, Harizuka S, Kumar S. Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques. *Emotion*. 2002; 2:75–84. [PubMed: 12899367]
- Ekman P, Sorenson ER, Friesen WV. Pancultural elements in facial displays of emotions. *Science*. 1969; 164:86–88. [PubMed: 5773719]
- Elfenbein HA, Foo MD, White J, Tan HH, Aik VC. Reading your counterpart: The benefit of emotion recognition accuracy for effectiveness in negotiation. *Journal of Nonverbal Behavior*. 2007; 31:205–223.
- Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*. 2007; 39:175–191.
- Feifel D, Macdonald K, Nguyen A, Cobb P, Warlan H, Galangue B, et al. Hadley A. Adjunctive intranasal oxytocin reduces symptoms in schizophrenia patients. *Biological Psychiatry*. 2010; 68:678–680. [PubMed: 20615494]

- Fertuck EA, Jekal A, Song I, Wyman B, Morris MC, Wilson ST, et al. Stanley B. Enhanced “Reading the Mind in the Eyes” in borderline personality disorder compared to healthy controls. *Psychological Medicine*. 2009; 39:1979–1988. [PubMed: 19460187]
- Gallese V, Keysers C, Rizzolatti G. A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*. 2004; 8:396–403. [PubMed: 15350240]
- Gallese V, Rochat M, Cossu G, Sinigaglia C. Motor cognition and its role in the phylogeny and ontogeny of action understanding. *Developmental Psychology*. 2009; 45:103–113. [PubMed: 19209994]
- Galobardes B, Shaw M, Lawlor DA, Lynch JW, Smith GD. Indicators of socioeconomic position (part 1). *Journal of Epidemiology and Community Health*. 2006; 60:7–12.
- Goding DC, Johnson M, Peterman JS. Schizotypy and altered digit ratios: A second look. *Psychiatry Research*. 2010; 178:73–78. [PubMed: 20471104]
- Golan O, Baron-Cohen S. Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia. *Development and Psychopathology*. 2006; 18:591–617. [PubMed: 16600069]
- Golan O, Baron-Cohen S, Hill J. The Cambridge Mindreading (CAM) Face Voice Battery: Testing complex emotion recognition in adults with and without Asperger syndrome. *Journal of Autism and Developmental Disorders*. 2006; 36:169–183. [PubMed: 16477515]
- Green MF, Olivier B, Crawley JN, Penn DL, Silverstein S. Social cognition in schizophrenia: Recommendations from the measurement and treatment research to improve cognition in schizophrenia new approaches conference. *Schizophrenia Bulletin*. 2005; 31:882–887. [PubMed: 16135561]
- Gregory C, Lough S, Stone V, Erzinclioglu S, Martin L, Baron-Cohen S, Hodges JR. Theory of mind in patients with frontal variant frontotemporal dementia and Alzheimer’s disease: Theoretical and practical implications. *Brain*. 2002; 125:752–764. [PubMed: 11912109]
- Guastella AJ, Einfeld SL, Gray KM, Rinehart NJ, Tonge BJ, Lambert TJ, Hickie B. Intranasal oxytocin improves emotion recognition for youth with autism spectrum disorders. *Biological Psychiatry*. 2010; 67:692–694. [PubMed: 19897177]
- Gunther Moor B, Op de Macks ZA, Güroglu B, Rombouts SA, Van der Molen MW, Crone EA. Neurodevelopmental changes of reading the mind in the eyes. *Social Cognitive and Affective Neuroscience*. 2012; 7:44–52. [PubMed: 21515640]
- Hall J, Harris JM, Sprengelmeyer R, Sprengelmeyer A, Young AW, Santos IM, et al. Lawrie SM. Social cognition and face processing in schizophrenia. *British Journal of Psychiatry*. 2004; 185:169–170. [PubMed: 15286070]
- Hallerbäck MU, Lugnegård T, Hjärthag F, Gillberg C. The Reading the Mind in the Eyes Test: Test-retest reliability of a Swedish version. *Cognitive Neuropsychiatry*. 2009; 14:127–143. [PubMed: 19370436]
- Happé FG. An advanced test of theory of mind: Understanding of story characters’ thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*. 1994; 24:129–154. [PubMed: 8040158]
- Harkness KL, Jacobson JA, Duong D, Sabbagh MA. Mental state decoding in past major depression: Effect of sad versus happy mood induction. *Cognition and Emotion*. 2010; 24:497–513.
- Harkness KL, Sabbagh MA, Jacobson JA, Chowdrey NK, Chen T. Enhanced accuracy of mental state decoding in dysphoric college students. *Cognition and Emotion*. 2005; 19:999–1025.
- Harrison A, Tchanturia K, Treasure J. Attentional bias, emotion recognition, and emotion regulation in anorexia: State or trait? *Biological Psychiatry*. 2010; 68:755–761. [PubMed: 20591417]
- Havet-Thomassin V, Allain P, Etcharry-Bouyx F, Le Gall D. What about theory of mind after severe brain injury? *Brain Injury*. 2006; 20:83–91. [PubMed: 16403703]
- Hu L, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*. 1999; 6:1–55.
- Ihnen GH, Penn DL, Corrigan PW, Martin J. Social perception and social skill in schizophrenia. *Psychiatry Research*. 1998; 80:275–286. [PubMed: 9796943]

- Irani F, Platek SM, Panyavin IS, Calkins ME, Kohler C, Siegel SJ, et al. Gur RC. Self-face recognition and theory of mind in patients with schizophrenia and first-degree relatives. *Schizophrenia Research*. 2006; 88:151–160. [PubMed: 16979876]
- Izard, CE. *The face of emotion*. New York, NY: Appleton-Century-Crofts; 1971.
- Kee KS, Kern RS, Green MF. Perception of emotion and neurocognitive functioning in schizophrenia: What's the link? *Psychiatry Research*. 1998; 81:57–65. [PubMed: 9829651]
- Kelemen O, Keri S, Must A, Benedek G, Janka Z. No evidence for impaired “theory of mind” in unaffected first-degree relatives of schizophrenia patients. *Acta Psychiatrica Scandinavica*. 2004; 110:146–149. [PubMed: 15233715]
- Kenyon M, Samarawickrema N, Dejong H, Van den Eynde F, Startup H, Lavender A, Goodman-Smith E, Schmidt U. Theory of mind in bulimia nervosa. *International Journal of Eating Disorders*. 2012; 45:377–384. [PubMed: 22212956]
- Kerr SL, Neale JM. Emotion perception in schizophrenia: Specific deficit or further evidence of generalized poor performance? *Journal of Abnormal Psychology*. 1993; 102:312–318. [PubMed: 8315144]
- Kettle JW, O'Brien-Simpson L, Allen NB. Impaired theory of mind in first-episode schizophrenia: Comparison with community, university and depressed controls. *Schizophrenia Research*. 2008; 99:96–102. [PubMed: 18155447]
- Köther U, Veckenstedt R, Vitzthum F, Roesch-Ely D, Pfueller U, Scheu F, Moritz S. “Don't give me that look”—overconfidence in false mental state perception in schizophrenia. *Psychiatry Research*. 2012; 196:1–8. [PubMed: 22482796]
- Kunihira Y, Senju A, Dairoku H, Wakabayashi A, Hasegawa T. “Autistic” traits in non-autistic Japanese populations: Relationships with personality traits and cognitive ability. *Journal of Autism and Developmental Disorders*. 2006; 36:553–566. [PubMed: 16602034]
- Lane RD, Merikangas KR, Schwartz GE, Huang SS, Prusoff BA. Inverse relationship between defensiveness and lifetime prevalence of psychiatric disorder. *American Journal of Psychiatry*. 1990; 147:573–578. [PubMed: 2327485]
- Larøi F, Fonteneau B, Mourad H, Raballo A. Basic emotion recognition and psychopathology in schizophrenia. *Journal of Nervous and Mental Disease*. 2010; 198:79–81. [PubMed: 20061875]
- Lawrence EJ, Shaw P, Baker D, Baron-Cohen S, David AS. Measuring empathy: Reliability and validity of the empathy quotient. *Psychological Medicine*. 2004; 34:911–919. [PubMed: 15500311]
- Lehmann, EL. *Nonparametrics. Statistical methods based on ranks*. San Francisco: Holden-Day; 1975.
- Leppänen JM, Hietanen JK. Emotion recognition and social adjustment in school-aged girls and boys. *Scandinavian Journal of Psychology*. 2001; 42:429–435. [PubMed: 11771812]
- Lombardo MV, Barnes JL, Wheelwright SJ, Baron-Cohen S. Self-referential cognition and empathy in autism. *PLoS ONE*. 2007; 2:e883. [PubMed: 17849012]
- Lorenzo-Seva U, Ferrando PJ. FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, Instruments and Computers*. 2006; 38:88–91.
- Machado-de-Sousa JP, Arrais KC, Alves NT, Chagas MH, de Meneses-Gaya C, Crippa JA, Hallak JE. Facial affect processing in social anxiety: Tasks and stimuli. *Journal of Neuroscience Methods*. 2010; 193:1–6. [PubMed: 20800619]
- Maggini C, Raballo A. Alexithymia and schizophrenic psychopathology. *Acta Bio Medica Ateneo Parmense*. 2004; 75:40–49.
- Matsumoto D, Olide A, Schug J, Willingham B, Callan M. Cross-cultural judgments of spontaneous facial expressions of emotion. *Journal of Nonverbal Behavior*. 2009; 33:213–238.
- Matsumoto D, Willingham B, Olide A. Sequential dynamics of culturally moderated facial expressions of emotion. *Psychological Science*. 2009; 20:1269–1275. [PubMed: 19754526]
- Miotto P, De Coppi M, Frezza M, Rossi M, Preti A. Social desirability and eating disorders: A community study of an Italian school-aged sample. *Acta Psychiatrica Scandinavica*. 2002; 105:372–377. [PubMed: 11942944]
- Miotto P, Preti A. Suicide ideation and social desirability among school-aged young people. *Journal of Adolescence*. 2008; 31:519–533. [PubMed: 17868799]

- Müller J, Buhner M, Ellgring H. Is there a reliable factorial structure in the 20-item Toronto Alexithymia Scale? A comparison of factor models in clinical and normal adult samples. *Journal of Psychosomatic Research*. 2003; 55:561–568. [PubMed: 14642988]
- Muller F, Simion A, Reviriego E, Galera C, Mazaux JM, Barat M, Joseph PA. Exploring theory of mind after severe traumatic brain injury. *Cortex*. 2010; 46:1088–1099. [PubMed: 19828142]
- Premack D, Woodruff G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*. 1978; 4:515–526.
- Preti A, Vellante M, Baron-Cohen S, Zucca G, Petretto DR, Masala C. The Empathy Quotient: A cross-cultural comparison of the Italian version. *Cognitive Neuropsychiatry*. 2011; 16:50–70. [PubMed: 20737328]
- Ramanaiah NV, Schill T, Leung LS. A test of the hypothesis about the two dimensional nature of the Marlowe-Crowne Social Desirability Scale. *Journal of Research in Personality*. 1977; 11:251–259.
- Rutherford MD, Baron-Cohen S, Wheelwright S. Reading the mind in the voice: A study with normal adults and adults with Asperger syndrome and high functioning autism. *Journal of Autism and Developmental Disorders*. 2002; 32:189–194. [PubMed: 12108620]
- Ryu V, An SK, Jo HH, Cho HS. Decreased P3 amplitudes elicited by negative facial emotion in manic patients: Selective deficits in emotional processing. *Neuroscience Letters*. 2010; 481:92–96. [PubMed: 20599588]
- Samson D. Reading other people's mind: Insights from neuropsychology. *Journal of Neuropsychology*. 2009; 3:3–16. [PubMed: 19338713]
- Schilling L, Wingenfeld K, Löwe B, Moritz S, Terfehr K, Köther U, Spitzer C. Normal mind-reading capacity but higher response confidence in borderline personality disorder patients. *Psychiatry and Clinical Neurosciences*. 2012; 66:322–327. [PubMed: 22624737]
- Schimansky J, Rössler W, Haker H. The influence of social cognition on ego disturbances in patients with schizophrenia. *Psychopathology*. 2012; 45:117–125. [PubMed: 22310555]
- Scrimin S, Moscardino U, Capello F, Altoè G, Axia G. Recognition of facial expressions of mixed emotions in school-age children exposed to terrorism. *Developmental Psychology*. 2009; 45:1341–1352. [PubMed: 19702396]
- Serafin M, Surian F. Il Test degli Occhi: uno strumento per valutare la "teoria della mente". *Giornale Italiano di Psicologia*. 2004; 4:839–860.
- Sijtsma K. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*. 2009; 74:107–120. [PubMed: 20037639]
- Smeets T, Dziobek I, Wolf OT. Social cognition under stress: Differential effects of stress-induced cortisol elevations in healthy young men and women. *Hormones and Behavior*. 2009; 55:507–513. [PubMed: 19470361]
- Stanford AD, Messinger J, Malaspina D, Corcoran CM. Theory of Mind in patients at clinical high risk for psychosis. *Schizophrenia Research*. 2011; 131:11–17. [PubMed: 21757324]
- Szanto K, Dombrovski AY, Sahakian BJ, Mulsant BH, Houck PR, Reynolds CF III, Clark L. Social emotion recognition, social functioning, and attempted suicide in late-life depression. *American Journal of Geriatric Psychiatry*. 2012; 20:257–265. [PubMed: 22354116]
- Tanaka, JS. Multifaceted conceptions of fit in structural equation models. In: Bollen, KA, Long, JS, editors. *Testing structural equation models*. Bollen, KA, Long, JS, editors. Newbury Park, CA: Sage; 1993. 10–39.
- Taylor GJ, Bagby RM, Parker JD. The 20-item Toronto Alexithymia Scale, IV. Reliability and factorial validity in different languages and cultures. *Journal of Psychosomatic Research*. 2003; 55:277–283. [PubMed: 12932803]
- Taylor GJ, Bagby RM, Ryan DP, Parker JDA, Doody KF, Keefe P. Criterion validity of the Toronto Alexithymia Scale. *Psychosomatic Medicine*. 1988; 50:500–509. [PubMed: 3186893]
- Tottenham N, Tanaka J, Leon AC, McCarry T, Nurse M, Hare TA, et al. Nelson CA. The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research*. 2009; 168:242–249. [PubMed: 19564050]
- Tsuruya N, Kobayakawa M, Kawamura M. Is "reading mind in the eyes" impaired in Parkinson's disease? *Parkinsonism and Related Disorders*. 2011; 17:246–248. [PubMed: 20889365]

- Valla JM, Ganzel BL, Yoder KJ, Chen GM, Lyman LT, Sidari AP, et al. Belmonte MK. More than maths and mindreading: Sex differences in empathizing/systemizing covariance. *Autism Research*. 2010; 3:174–184. [PubMed: 20589713]
- Von Horn A, Bäckman L, Davidsson T, Hansen S. Empathizing, systemizing and finger length ratio in a Swedish sample. *Scandinavian Journal of Psychology*. 2010; 51:31–37. [PubMed: 19392945]
- Voracek M, Dressler SG. Lack of correlation between digit ratio (2D:4D) and Baron-Cohen's "Reading the Mind in the Eyes" test, empathy, systemising, and autism-spectrum quotients in a general population sample. *Personality and Individual Differences*. 2006; 41:1481–1491.
- White S, Hill E, Happé F, Frith U. Revisiting the strange stories: Revealing mentalizing impairments in autism. *Child Development*. 2009; 80:1097–1117. [PubMed: 19630896]
- Yildirim EA, Kasar M, Güdük M, Ates E, Küçükparlak I, Özalmete EO. Investigation of the reliability of the "reading the mind in the eyes test" in a Turkish population. *Türk Psikiyatri Dergisi*. 2011; 22:177–186. [PubMed: 21870307]
- Zinbarg RE, Yovel I, Revelle W, McDonald RP. Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω . *Applied Psychological Measurement*. 2006; 30:121–144.

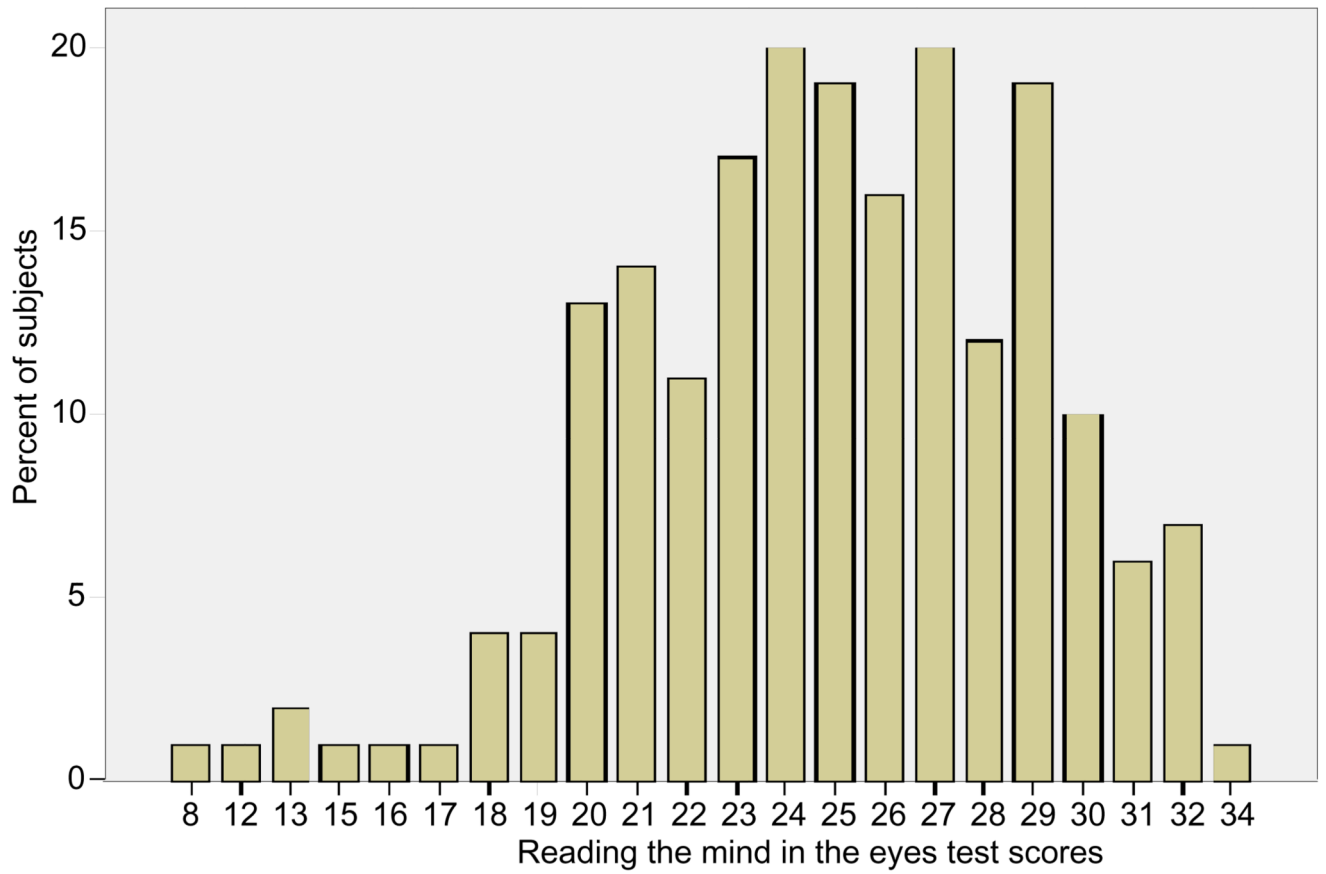


Figure 1. Distribution of test scores, 36 items ($n = 200$).

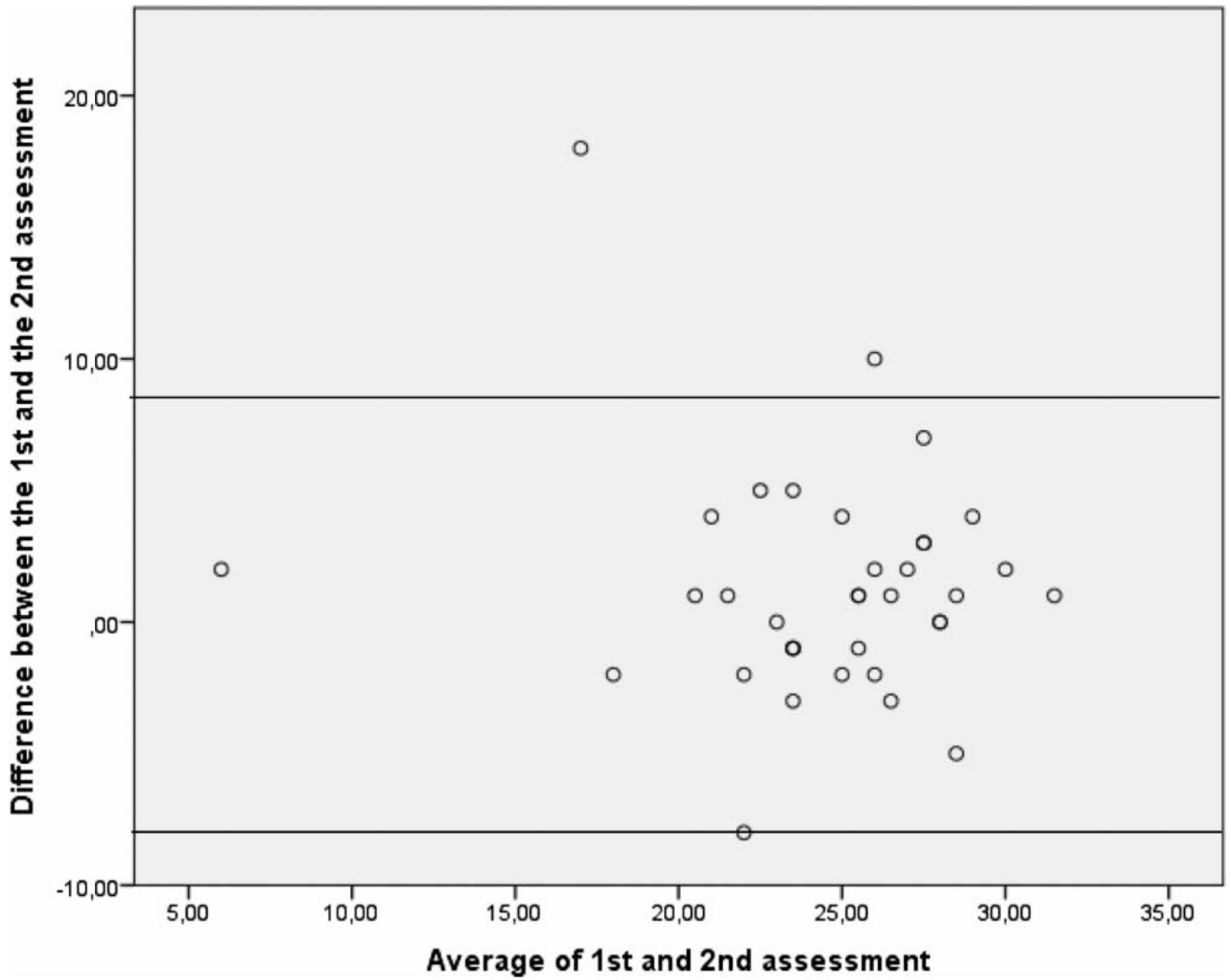


Figure 2. Bland-Altman plot of the Eyes Test test-retest assessment ($n = 36$).

Table 1
Psychometric characteristics of the Eyes test in past studies

Study	Country	Sample	No. of items	Reliability	Mean score	Sex differences	Correlation with measures of empathy	Association with IQ
Ahmed & Miller (2011)	Athens, USA	123 healthy people 70 females, 53 males age: 19.0 ± 1.3	36	Unreported	27.3 ± 3.7	No differences by gender	Untested	Verbal IQ related to Eyes test scores
Alaerts et al. (2011)	Heverlee, Belgium	37 healthy people 22 females, 15 males age: 24.6 ± 5.1	36	Unreported	Unreported	Statistically significant advantage for females	Ad hoc emotion recognition task: $r = .38$	Untested
Ali et al. (2010)	London, UK	112 undergraduate students age: 20.7 ± 5.8	36	Unreported	Unreported	Unreported	EQ: $r = .44$	Untested
Baron-Cohen, Wheelwright, Hill, et al. (2001)	Cambridge, UK	103 undergraduate students 50 females, 53 males age: 20.8 ± 0.8	36	Unreported	Men: 27.3 (3.7) Women: 28.6 (3.2)	No differences by gender	Untested	No links with IQ
Carroll & Chiew (2006)	York, UK	48 university students 24 females, 24 males age: 21	36	Unreported	Men: 24.8 (4.6) Women: 27.8 (3.6)	Cohen's $d = 0.71$ favouring females	EQ ($n = 20$): $r = .36$	No impact of IQ on the links between Eyes test and EQ
Chapman et al. (2006)	Cambridge, UK	76 children 37 girls, 39 boys age: 6 to 9 years old	28	Unreported	Boys: 15.2 (3.5) Girls: 16.3 (3.3)	No differences by gender	EQ-c: $r = .56$	No links with IQ
Cook & Saucier (2010)	Lethbridge, Canada	88 undergraduate students age: 22.2 ± 5.6	36	Unreported	Men: 26.1 (3.7) Women: 27.2 (4.1)	No differences by gender	EQ: $r = .31$	Untested
Dehning et al. (2012)	Jimma city, Ethiopia	237 medical students 30 females, 207 males age: 21.4 ± 2.7	36	Cronbach's alpha: .70	First year: Men: 14.2 (4.7) Women: 16.3 (4.8) Final year: Men: 17.2 (4.3) Women: 21.2 (4.0)	Cohen's $d = 0.45$ and 0.94, respectively in first and final year, favouring females	Untested	Untested
Golan & Baron-Cohen (2006)	Cambridge, UK	39 people with normal IQ (26 with autistic traits) 7 females, 32 males age: 25.1 ± 8.4	36	Unreported	Unreported	Unreported	Untested	Verbal IQ related to Eyes test scores
Hallenäck et al. (2009)	Karlstad, Sweden	158 undergraduate students 83 females, 75 males age: 23.9 ± 3.3	28	Stable at retest according to the Bland-Altman plot	Men: 18.5 (1.9) Women: 19.2 (2.3)	Cohen's $d = 0.33$ favouring females	Untested	Untested

Study	Country	Sample	No. of items	Reliability	Mean score	Sex differences	Correlation with measures of empathy	
							Association with IQ	Association with EQ
Harkness et al. (2010)	Waterloo, Canada	93 college students 76 females, 17 males age: 18.9 ± 1.4	36/28	Cronbach's alpha: .58 in the 36 items version, .63 in the 28 items version	Men: 17.6 (3.4) Women: 19.0 (3.6)	Nonsignificant advantage for the women	Untested	Untested
Kelemen et al. (2004)	Keckskemet, Hungary	40 controls 12 females, 28 males age: 41.2 ± 15.4	36	Unreported	27.8 (5.0)	Unreported	Untested	No links with IQ
Kenyon et al. (2012)	London, UK	57 controls age: 24.0 ± 2.6	36	Unreported	28.3 (3.2)	Unreported	Untested	IQ $r = .28$ in controls
Kettle et al. (2008)	Parkville, Australia	43 healthy controls 29 females, 14 males age: 19.2 ± 0.6	36	Unreported	Cannot be derived	No differences by gender	Untested	Untested
Kunihira et al. (2006)	Tokyo, Japan	40 undergraduate students age: cannot be derived	32	Unreported	Cannot be derived	No differences by gender	Untested	Untested
Lawrence, Shaw, Baker, Baron-Cohen & David (2004)	London, UK	53 adult volunteers 28 females, 25 males age: 32.5 ± 10.9	36	Unreported	Unreported	Unreported	EQ social skills factor: $r = .27$	Untested
Muller et al. (2010)	Bordeaux, France	15 controls age: 37.0 ± 12.5	20	Unreported	13.0 (1.9)	Unreported	No links with Davis' IRI	Untested
Serafin & Surian (2004)	North-east Italy	75 healthy people 41 females, 34 males age: 26.7 ± 5.4	36	Guttman's split-half: .77	Men: 24.3 (3.7) Women: 25.5 (4.0)	No differences by gender	Untested	Untested
Smeets, Dziobek, & Wolf (2009)	???, Germany	64 undergraduate students 32 females, 32 males age: 25.9 ± 4.3	36	Unreported	Men: 28.4 (1.1) Women: 28.4 (1.1)	No differences by gender	Untested	Untested
Stanford, Messinger, Malaspina, & Corcoran (2011)	New York, USA	115 subjects 39 controls, 13 with schizophrenia, 63 with high risk of psychosis 38 females, 77 males	36	Unreported	Young people ($n = 24$): 25.7 (4.9)	No differences by gender	Untested	IQ related to Eyes test scores
Valla et al. (2010)	New York, USA	144 undergraduate students 79 females, 65 males age: 20.2 ± 1.8	36	Unreported	Men: 27.1 (3.5) Women: 27.3 (3.2)	No differences by gender	Untested	Untested

Study	Country	Sample	No. of items	Reliability	Mean score	Sex differences	Correlation with measures of empathy	Association with IQ
Voracek & Dressler (2006)	Vienna, Austria	423 healthy people age: 29.7 ± 11.4	36	Cronbach's alpha: .63 in men, .60 in women	Men: 22.3 (4.5) Women: 23.3 (4.2)	Cohen's $d = 0.22$ favouring females	EQ: $r = .23$	Untested
Yildirim et al. (2011)	Bakirkoy, Turkey	130 healthy people 72 females, 45 males age: 33.3 ± 8.9	36	Test-retest ICC: .65 (95% CI: .49 to .77)	Men: 23.2 (3.5) Women: 25.2 (3.1)	Cohen's $d = 0.61$ favouring females	Untested	Untested

Table 2
General characteristics of the sample ($n = 200$)

Sociodemographic group	N (%)	Eyes test Mean (SD)
Gender		
Male	92 (46%)	24.1 (4.7)
Female	108 (54%)	25.5 (3.5)*
Age		
18–24 (%)	115 (57%)	24.5 (4.6)
25–32 (%)	85 (43%)	25.3 (3.5)
Highest level of parental education		
Lower than high school diploma	89 (44%)	24.6 (3.8)
High school diploma	82 (41%)	24.9 (4.4)
College graduate or higher	29 (15%)	25.1 (4.6)

* t -test $p < .05$.

Table 3
Reading the Mind in the Eyes test—Distribution of responses in percentages*

Item	Answer A	Answer B	Answer C	Answer D	Corrected		Factor loading	
					M	F		
1	69.5 Comforting	24.5 Bored	3.5 Irritated	2.5 Irritated	2.5	66.3	72.2	.168
2	Terrified	18.0 Upset	18.5 Annoyed	18.5 Arrogant	7.5	51.1	60.2	.242
3	Joking	0.0 Flustered	7.5 Convinced	27.5 Desire	65.0	60.9	68.5	.001
4	Joking	4.0 Insisting	27.5 Relaxed	27.5 Amused	3.0	65.2	65.7	-.031
5	Irritated	2.5 Sarcastic	13.5 Friendly	0.0 Worried	84.0	78.3	88.9	.148
6	Aghast	7.5 Fantasing	69.0 Alarmed	2.5 Impatient	21.0	60.9	75.9	.315
7	Apologetic	19.0 Friendly	25.5 Dispirited	10.0 Uneasy	42.5	39.1	45.4	.024
8	Despondent	67.0 Relieved	5.0 Excited	22.0 Shy	6.0	67.4	66.7	.164
9	Annoyed	5.0 Hostile	3.5 Preoccupied	90.5 Horrified	1.0	89.1	91.7	.254
10	Cautious	63.5 Insisting	26.5 Aghast	5.0 Bored	5.0	59.8	66.7	.050
11	Terrified	6.0 Amused	5.5 Flirtatious	17.5 Regretful	71.0	69.6	72.2	.076
12	Indifferent	19.0 Embarrassed	8.5 Dispirited	1.0 Sceptical	71.5	73.9	69.4	-.058
13	Decisive	14.5 Anticipating	63.5 Shy	21.0 Threatening	1.0	54.3	71.3	.203
14	Irritated	10.5 Disappointed	6.5 Accusing	80.0 Depressed	3.0	78.3	81.5	.317
15	Contemplative	83.0 Flustered	2.5 Amused	2.5 Encouraging	12.0	78.3	87.0	.511
16	Irritated	2.0 Thoughtful	76.0 Sympathetic	19.5 Encouraging	2.5	67.4	83.3	.409
17	Doubtful	54.0 Affectionate	32.0 Aghast	9.5 Playful	4.5	58.7	50.0	.049
18	Decisive	92.0 Amused	2.5 Bored	2.5 Aghast	3.0	91.3	92.6	.231
19	Arrogant	4.0 Graeful	34.0 Tentative	Sarcastic	9.5	53.3	51.9	.104
20	Dominant	18.5 Friendly	73.5 Horrified	1.5 Guilty	6.5	68.5	77.8	.178
21	Embarrassed	15.5 Fantasing	73.0 Panicked	2.5 Confused	9.0	67.4	77.8	.313
22	Preoccupied	90.5 Grateful	0.5 Imploring	4.5 Insisting	4.5	88.0	92.6	.317
23	Contented	5.5 Apologetic	9.5 Curious	22.5 Defiant	62.5	62.0	63.0	.152
24	Pensive	58.5 Irritated	9.0 Hostile	7.5 Excited	25.0	56.5	60.2	.096
25	Panicked	6.5 Incredulous	8.5 Interested	67.0 Despondent	18.0	65.2	68.5	.352
26	Alarmed	6.5 Shy	5.0 Anxious	12.0 Hostile	76.5	67.4	84.3	.333
27	Joking	2.5 Cautious	63.0 Reassuring	18.5 Arrogant	16.0	57.6	67.6	.361

<i>Item</i>	<i>Answer A</i>	<i>Answer B</i>	<i>Answer C</i>	<i>Answer D</i>	<i>Corrected</i>			<i>Factor loading</i>
					<i>M</i>	<i>F</i>	<i>F</i>	
28	70.0 Interested	Joking 3.5	Contended 22.0	Affectionate 4.5	69.6	70.4	70.4	.404
29	9.5 Impatient	Aghast 10.0	Reflective 66.5	Irritated 14.0	71.7	62.0	62.0	.140
30	4.5 Grateful	Flirtatious 91.0	Disappointed 2.0	Hostile 2.5	87.0	94.4	94.4	.264
31	Ashamed 7.5	Confident 66.5	Dispirited 18.5	Joking 7.5	63.0	69.4	69.4	.315
32	Serious 73.0	Ashamed 3.0	Alarmed 10.5	Bewildered 13.5	70.7	75.0	75.0	.394
33	Embarrassed 9.5	Guilty 23.0	Concerned 54.0	Fantatising 13.5	55.4	52.8	52.8	.171
34	Aghast 2.0	Baffled 21.5	Terrified 5.5	Distrustful 71.0	71.7	70.4	70.4	.139
35	Puzzled 34.0	Nervous 36.5	Contemplative 13.5	Insisting 16.0	44.6	29.6	29.6	-.035
36	Ashamed 1.0	Nervous 7.0	Indecisive 15.5	Suspicious 76.5	77.2	75.9	75.9	.319

* Correctly identified descriptors are marked in bold. Factor loading refers to the unidimensional model.

Table 4
Participants' scores on the self-report scales (top) and relations between Eyes test scores and the other measures (bottom)

Cronbach's α	Males (n = 92)				Females (n = 108)				Total sample (n = 200)				
	Minimum	Maximum	Mean (SD)	IQR	Minimum	Maximum	Mean (SD)	IQR	Minimum	Maximum	Mean (SD)	IQR	
Eyes test	8	32	24.1 (4.7)	15	34	25.5 (3.5)	25	5	15	67	45.3 (9.1)	45	13
EQ	25	64	41.8 (8.7)	27	72	43.1 (9.0)	43	12	27	72	43.7 (9.1)	43	12
TAS	25	71	44.4 (9.1)	22	28	16.3 (4.5)	16	7	22	28	16.5 (4.8)	16	7
SDS	3	29	16.6 (5.2)	5									

<i>Pearson's correlation coefficient</i>						
	Eyes test	EQ	TAS	Eyes test	EQ	TAS
EQ	-.039			-.070		
TAS	-.220*	-.358**		.019	-.455**	
SDS	.079	.302*	-.321*	.013	.410**	-.325*

	Eyes test	EQ	TAS
EQ			
TAS			
SDS			

SD=standard deviation; IQR = interquartile range. * $p < .01$, ** $p < .0001$.