Validation and psychometric properties of the "Reading the Mind in the Eyes Test"

Renaud F. Cohen, Anna Maria Berardi, Pascal Le Vaou, Jean-Pierre Kahn

Validation and psychometric properties of the "Reading the Mind in the Eyes Test"

Renaud F. Cohen[1, 2,3] ; Anna Maria Berardi [2]; Pascal Le Vaou [4] ; Jean-Pierre Kahn [1, 2,3]

1 : Centre Psychothérapique de Nancy, 1, rue du Docteur Archambault, BP 11010, 54 521 LAXOU cedex. FRANCE.

2 : Université de Lorraine, FRANCE.

3 : FACE-BD network, fondation FondaMental

4: CHR Metz-Thionville

Renaud F. Cohen

Mail : renaudcohen@hotmail.fr

Postal address :

1, rue du Docteur Archambault, BP 11010, 54 521 LAXOU cedex. FRANCE

Phone number: +33675870510

Professional phone number: +33383926844

Conflicts of interest: none

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Abstract

The French validation and psychometric investigation of the "Reading the Mind in the Eyes Test" (or "Eyes Test"; Baron-Cohen et al., 2001) was carried out on 661 French participants. Participants completed the Eyes Test, Facial Emotional Recognition test and Mill-Hill Vocabulary. This study is, to the best of our knowledge, the largest study assessing the psychometrics characteristics of the Eyes Test, allowing a Bayesian Item Response Model analysis and the study of its presumed unidimensionality. A subsample of 71 participants completed the Reading the Mind in the Eyes Test twice. The 3-PL model of Item Response Theory with Bayesian estimation was used to find the items with significant discriminant alpha coefficients. Parallel Analysis and Confirmatory Factor Analysis were used to assess the unidimensionality of the Eyes Test. Thirty one items were included in the final version of the French Eyes Test. Results suggest that the French Eyes Test with 31 items has good convergent and discriminant validities and that it fits a one-dimensional model ; moreover, the test is stable across time. However, the 3-PL Bayesian Item Response Theory Model fit suggests a high level of correct guessing. The Eyes Test measures a unique ability, namely affective theory of mind.

Keywords: Reading the Mind in the Eyes Test, Theory of mind, psychometric properties, Bayesian item response theory model, social cognition, French validation.

1. INTRODUCTION

Theory of Mind (ToM) is a reflection of social cognition and includes all the abilities involved in the relationships between humans and their social environment. An NIMH Workshop on Social Cognition in Schizophrenia defined social cognition as "the mental operations that underlie social interactions, including perceiving, interpreting, and generating responses to the intentions, dispositions, and behaviors of others" (Pinkham et al., 2014). The affective ToM corresponds to the ability to infer the roots of emotions identified in other people (Shamay-Tsoory & Aharon-Peretz, 2007). This ability is frequently assessed with the 'Reading the Mind in the Eyes Test' (Baron-Cohen et al., 2001) or validated translations of this test (Hallerbäck, Lugnegard, Hjärthag, & Gillberg, 2009; Sanvicente-Vieira et al., 2013; Vellante et al., 2013).

In the Eyes Test, participants are presented with a series of 36 photographs of the eye-region of different actors and actresses, and are asked to choose which of four words best describes what the person in the photograph is thinking or feeling (Baron-Cohen et al., 1997, 2001). The target words and distractors are always complex mental states, like "reflective" or "irritated" (see the revised version; Baron-Cohen et al., 2001). Therefore, this test assesses affective ToM (Shamay-Tsoory et al., 2007), as it involves, for every item, the ascription of complex mental states (Baron-Cohen et al., 2001) and not of "basic emotions" (e.g. Ekman, 1992a, 1992b).

The Eyes Test's capacity to discriminate between subtle variations of ToM levels is exemplified by Domes et al. (2007), who showed that an intranasal administration of oxytocin improves the correct recognition of target words on the Eyes Test in healthy participants and by the evidence of a superior average performance in female compared to male participants on this test, a result that has been confirmed in a meta-analysis of the 'Eyes Test' (Kirkland, Peterson, Baker, Miller, & Pulos, 2013).

Therefore, it is one of the few tests that can assess high levels of performance in social cognition in healthy adults, where most other tests show strong ceiling effects. But, the hypotheses, that the test assesses a unique latent ability (i.e. attribution of complex emotion), deserves an empirical psychometric investigation. Therefore, the validity of the Eyes Test has been studied, but this does not imply a unique latent ability (unidimensionality of the test).

Our aim was also to validate the Eyes Test in French, because it is the most powerful test to assess ToM, and to demonstrate its psychometric properties, among which unidimensionality, validity and stability across time. To this purpose, a Bayesian 3-parameter logistic item response theory (IRT) model was fitted to compute all the coefficients associated with IRT, namely difficulty (beta), discriminant (alpha) and guessing (eta) parameters of the items (Lord, 1980). This article is the first to use Bayesian IRT analysis to study the psychometric properties of the Eyes Test (Fox, 2010). We were particularly interested in the evaluation of the alpha parameters (which have to be superior to 0) and the guessing parameters. Indeed, Johnston, Miles, & McKinlay (2008) found that a group of healthy subjects could infer the target word without the picture associated with it, suggesting high levels of correct guessing in the original test. The 3-PL model could be computed to obtained the "eta" paremeters, which reflect the items' associated levels of correct guessing. This parameter is particularly meaningful, indeed, it corresponds to the probability that a subject with very low levels of ability responds correctly to the item (DeMars, 2010).

One can expect the Eyes Test to correlate with the Facial Emotion Recognition Test (FERT), because even if the abilities assessed by the Eyes Test are wider than those assessed by FERT, in both tests facial emotional cues need to be interpreted. Therefore, a positive correlation between these two tasks of cognitive empathy was expected.

Demonstration of the unidimensionality of a test is a pre-requisite to the use of its total score. However, previous studies of the Eyes Test have generally neglected this psychometric

central element, apart from the studies carried out by Preti, Vellante, & Petretto (2017) and Vellante et al. (2013). The significance of this step in test validation has been pointed out by Ziegler & Hagemann (2015).

Vellante et al. (2013) used "Confirmatory Factor Analysis" (CFA) to test two models on the Italian Eyes Test. The first model was defined as unidimensional, i.e. all items loaded on a single latent dimension or factor. This model fitted well the data. The second model was defined by three dimensions and did not fit the data. The results supported the unidimensionality of the Italian Eyes Test. However, the method chosen did not take into account the binary nature of the variables. The second part of their analysis, reported in Preti et al. (2017), took into account the binary nature of the items with an IRT analysis, but did not analyze the guessing parameters.

The soundest methods to study unidimensionality are Confirmatory Factor Analysis (CFA) and Parallel Analysis (PA) (Cho, Li, & Bandalos, 2009; Ziegler & Hagemann, 2015). The CFA can help one to determine if one factor alone explains all responses to the test items (Brown, 2006). This method can be complemented with the optimal implementation of parallel analysis (Cho et al., 2009; Drasgow & Lissak, 1983; Lorenzo-Seva & Ferrando, 2013).

Therefore, the aim of this study was to assess the reliability (including the unidimensionality), and the convergent and discriminant validities of the French Eyes Test. We also wanted to study the fit of a 3-PL Bayesian IRT model, to select the items with non-zero alpha parameters and to demonstrate the importance of guessing in this test (by studying guessing parameters). These analyses have never been carried out on the French Eyes test, on its original English version, or in any of its validated translations.

2. METHODS

2.1.Participants

The sample was composed of 661 healthy French-speaking university students and their relatives, a random subsample of which took the Eyes Test twice (n =71). Participants were recruited on a voluntary basis. Five hundred twenty five women and 129 men take part to the study with nine subjects with unavailable gender information. The mean age was 23.1 years (±7.25). A psychologist (RFC) explained and completed the assessments. Participants were recruited from a University campus in France and later by the snow-ball sampling method.

All of the procedures were approved by the ethical committee of the University Medical Center of Nancy (CHU de Nancy, Université de Lorraine, France).

.A random subsample of 71 participants took part in the test twice at least one week apart (mean = 13.6 ± 9.9 days).

2.2.Tests administered

Participants were administered a questionnaire to record their demographic characteristics and health status. A pre-validated version of the French Eyes Test with 35 items (see below for its characteristics), was administered to all participants. The performances on all items were analyzed, including the practice item of the English version.

Simple emotion recognition was assessed with the *Facial Emotion Recognition Test* (FERT), a 35 item set of faces from the Ekman and Friesen (1975)'s set of pictures of facial affect; it includes five items for each basic emotion (fear, sadness, disgust, anger, surprise and happiness) and 5 items for neutral pictures. This classic task requires that subjects identify the emotional display of the face of an actor. The total score was used to obtain an index of global recognition of basic emotions. FERT was completed by 614 participants (participants with missing data were excluded).

Verbal intelligence was assessed with the Mill-Hill Vocabulary Test (Raven and Deltour, 1998) on a subsample of participants (n = 102 with complete data) to test for the separate effects of global verbal intelligence and of affective ToM. The information collected was anonymous.

2.3.Procedures

All participants were administered the items of the Eyes and Elman's tests on a video screen. For the Eyes test, participants responded by pressing one of four buttons (1,2,4,5; according to the position of the target word on the screen) on a computer, and, for the Ekman's test, they selected the target word with the mouse.

Mill-Hill was administered in paper form to a subsample of 102 subjects. The size of this sample is sufficient to detect modest positive associations with either Eyes or Ekman's tests.

The Eyes Test is highly vulnerable to cultural factors (Hallerbäck et al., 2009; Sanvicente-Vieira et al., 2013; Yıldırım et al., 2011), a problem that elicited difficulties in previous validations in other languages. In their original version of the Eyes Test, Baron-Cohen et al. (2001) did not know what the persons in the pictures were thinking or feeling; this was determined by subjective judgments using groups of eight judges (Baron-Cohen et al., 1997, 2001). In the current study, the English Eyes Test was first translated into French and back-translated. Second, the original procedure of Baron-Cohen et al. (2001) with eight judges was used to select reliable items. Five judges out of eight had to agree and no more than two judges had to choose the same distractor as the target. After successive assessments of this type, a 35-items test without any practice items was obtained. An item from the child version of the test, with new words (target and distractors), was added. We carried out the statistical analyses described below on this 35-items pre-validated French version of the Eyes Test to select the items to be included in the final version.

2.4.Statistical procedures

The statistical analyses were carried out with the software R version 3.2.0 (R Development Core Team, 2014), FACTOR 10 (Lorenzo-Seva & Ferrando, 2013) and JAGS software (Lunn, Spiegelhalter, Thomas, & Best, 2009; Plummer, 2003), interfaced with R with the R2jags package.

2.4.1. Statistical procedures to obtain the final version of the French Eyes Test

To select the items for the final French version of the Eyes Test, a three-step procedure was followed. First, a 3-parameter logistic model (Lord, Novick, & Birnbaum, 1968) with Bayesian estimation was fitted as described in Curtis et al (2010). JAGS software interfaced with R (R2jags package) was used on the full test (35 items). Markov Chain Monte Carlo Methods estimation were used to obtanin full posterior distributions of the parameters of the 3-PL IRT model with the script from Curtis (2010). A small modification of the script was included, allowing for a small negative alpha in the prior random distribution (left truncation at $\alpha = -0.5$ with a prior normal distribution of m =1 ; sd = 1) This analysis allowed to compute the alpha parameters to observe if the credible interval at 95 % included 0 for some of the parameters, whereas the majority of alpha are constrained to be positive. This procedure would indicate the non-discriminant items to be suppressed from further analysis.

We used a burn-in period of 10 000 iterations, then we launched 100 000 iterations for 3 chains (total iterations = 300 000), with a thinning interval of 10. Therefore, we used 30 000 iterations for each parameter.

To confirm that the chains converge, multiple methods were used (Kruschke, 2015). The convergence of the 3 chains for each item (mixing of the 3 chains) was inspected visually. The potential scale reduction factor was used and should tend towards 1 at convergence (see Gelman & Rubin, 1992). Another diagnostic is the Geweke diagnostic, which takes two nonoverlapping parts (usually the first 0.1 and the last 0.5 proportions) of the Markov chain

and compares the means of both parts by using a difference among the means to verify whether the two parts of the chain are from the same distribution (Geweke, 1991). This last index could be used with a unique chain, but the 3 chains were successively tested and only the statistics, which were significant in at least two of the chains for the same parameter, were considered as significant.

Second, a Confirmatory Factor Analysis (CFA, with the *lavaan* package: Rosseel, 2012) to the remaining 31 items was fitted with the diagonal Weight Least Square method with Mean and Variance adjustment (WLSMV), which is specifically efficient with dichotomous items (Beauducel & Herzberg, 2009), to study their unidimensionality. This analysis was confirmed with the optimal implementation of parallel analysis as described in Lorenzo-Seva & Ferrando (2013) with FACTOR 10 software.

The fit of the one-factor CFA model on the Eyes Test-31 was assessed with the following criteria: RMSEA (Root Mean Square Error of Approximation) <.05 (good fit of the model), CFI (comparative fit index) > .90, TLI (Tucker-Lewis index) > .90 (Hu & Bentler, 1998, 1999) and WRMR (weighted root-mean-square residual) < .90 (Yu, 2002). If these four criteria were fulfilled, the model was considered to fit well the data. The standard criteria were used for parallel analysis based on tetrachoric correlations (see Lorenzo-Seva & Ferrando, 2013).

The most consensual indices of reliability were computed to study the homogeneity of the Eyes Test-31: McDonald's omega coefficients (McDonald, 1999), using the method suggested by Gadermann, Guhn, and Zumbo (2012), who proposed tetrachoric correlations rather than Pearson correlations to compute reliability indices for binary data. Shapiro-Wilk tests were used to determine if the scores followed a normal distribution.

Thirdly, a 3-parameter logistic model with Bayesian estimation (Fox, 2010) was fitted again on the 31 remaining items. This analysis (using the same computing characteristics as

described earlier) allowed us to compute the posterior distribution of guessing parameters and to study their characteristics (median values and distributions) and therefore their potential utility.

2.4.2. Discriminant and convergent validities

To study convergent and discriminant validities, we computed two Spearman's correlations between the French Eyes-Test-31 and both FERT and Mill-Hill. A positive correlation between FERT and French Eyes-Test-31would support convergent validity, because they both assess emotion recognition, whereas a non significant correlation between Mill-Hill and French Eyes-Test-31would support discriminant validity (the relative independence of the Eyes Test from intellectual level). A positive partial correlation between French Eyes-Test-31and FERT, adjusted on Mill-Hill score, would discard the hypothesis that the relation between the French Eyes-Test-31and FERT is linked to non-specific factors like verbal or global intelligence.

2.4.3. Test-retest stability

The intraclass correlation coefficient (ICC) "agreement" was computed to compare the test-retest stability of performance on a subsample of 71 participants (Weir, 2005). A high level of this coefficient (>.70) indicates that the absolute score of the participants does not differ between the two occasions.

3. RESULTS

3.1.Selection of items for the French Eyes Test-31

The French pre-validated version of 35 items was the point of departure for the selection of items to be included in the final French Eyes Test.

A 3-PL Bayesian IRT model was fitted on the 35 items and the credible intervals of the discriminant coefficients alpha were examined. The credible intervals included 0 for four items. Therefore, 31 items remained for the CFA (cf. table 1).

Insert Table 1 here

3.2.One factor Confirmatory Factor Analysis

The fit for the 31-items of the final French Eyes Test was good for the DWLS estimation (followed by Robust estimation) : CFI = .950 (.861), TLI = .946 (.851), RMSEA = .012 (.015) CI = [.000, .018] < .05, SRMR = 0.081, WRMR = .979 ≈ .90. These fit indices are acceptable according to generally accepted criteria (McDonald & Ho, 2002; Yu, 2002), but the robust estimation is less satisfactory than the classic estimation. The one factor model is supported. The modification indices suggest that including error covariance between residuals of items 4 and 35 could increase the fit of the model. With this modification the fit is even better: CFI = .0.964 > .90 (.879), TLI = .962 > 90 (.870), RMSEA = .010 (.014) CI = [0.000, 0.017] < .05, SRMR= .080, WRMR = .965 ≈ .90.

3.3.Optimal implementation of Parallel Analysis

According to the chosen criteria, the parallel analysis confirmed the unidimensionality, with only one observed eigenvalue superior to the 95^{th} quantiles of random eigenvalues (results from FACTOR 10 software).

3.4.Statistical characteristics of the Eyes Test–31

The final test includes 17 pictures of the eyes-region of men and 14 similar pictures of women. The reliability coefficients, based on the tetrachoric correlation matrix, are ordinal omega ω = .79.

The Shapiro-Wilk test is significant (W = 0.954, p < .0001), suggesting that the distribution of scores does not follow a normal distribution (see figure 1). The confidence interval of the total score at 90 percent (from 5 to 95 quantile) is CI = [18 – 29].

Insert Figure 1 here

The total average of the means of each item on the 661 participants is .776 and the average of the standard deviations for each item is .389. The range of item-total corrected correlations in our sample was of 0.092 to 0.406 with a median of r_{pbis} = 0.202.

The properties of the items are presented in table 2, where one can see that the target words that obtained the lowest percent of correct responses (44.8 % for item 26 and 47.5 % for item 8) have a significant discriminant coefficients in the 3-PL model (their credible interval did not include 0). Therefore two items obtained a mean success score inferior to .50 and their distractors obtained more than 25 percent of total responses (Baron-Cohen et al.'s criteria, 2001); but they present good psychometrical characteristics in the 3-PL IRT model (discriminant coefficients superior to 0) and the CFA model.

Insert Table 2 here

3.5. Three-Parameters Logistic Bayesian Item Response Model Estimation of the Eyes Test-31

The 3 chains converge after the initial burn-in period of 10 000 iterations, as suggested by visual inspection showing mixing of the chains, with a potential scale reduction factor tending to 1 and Geweke diagnostics being non significant for each parameter in more than one chain. All alpha parameters' credible intervals are superior to zero (quantiles 2.5% of alpha parameters > 0), confirming that all items contribute to the measure of the unique latent dimension.

Insert Table 3 here

The guessing coefficients (eta) are particularly interesting (see table 3). The mean of the 31 eta coefficients is 0.504 [min = 0.154; max = 0.820] (estimated as the median of the values obtained at each iteration), which is substantially higher than 0.25. We could have hypothesed the .25 value from the fact that it is a multiple choice decision with 4 possible choices. Only two items obtained a computed median value inferior to .25 (item 8 and item 19). Twelve items out of 31 had 95% credible intervals which did not include .25 (strictly superior to .25). These results show a remarkably high level of guessing.

The high levels of guessing and the evidence for the good convergence of the model estimation indicate the usefulness of computing a 3-PL item response model for the French Eyes-Test-31.

3.6.Convergent and discriminant validities of the Eyes Test-31

The Spearman correlation between the Mill-Hill and the French Eyes-Test-31is not significant in the subsample with complete data on all three tests ($N = 102$; $r = .11$, $p > .10$). The correlation between FERT and Mill-Hill is statistically significant ($r = .26$; $p < .01$). Moreover, even after partialling out the Mill-Hill score, the partial correlation between the French Eyes-Test-31and FERT is still significant ($r = .22$; $p < .05$). Therefore, as expected, the French Eyes-Test-31 is associated with the FERT even when the global effect of intellectual level is statistically controlled. Finally, the results suggest a stronger correlation between FERT and Mill-Hill than between the French Eyes-Test-31and Mill-Hill.

3.7.Test-retest reliability

The ICC for agreement was .749, CI= [.62, .84] suggesting that the scores are stable across two evaluations.

4. DISCUSSION

The psychometric properties of the French Eyes Test-31 are similar to those of the original English version (Baron-Cohen et al., 2001). The results showed that a unique latent dimension could explain the pattern of performance on this test, namely the affective ToM. The results showed good reliability and validity of the Eyes Test-31 (test-retest reliability, convergent and discriminant validities).

4.1.Psychometric properties of the French Eyes Test-31

Given the criticisms with respect to Cronbach's alpha, a more consensual coefficient was computed, namely the "ordinal" version of the *omega coefficient* (Gadermann et al., 2012; McDonald, 1999) $\omega = .79$, which is satisfactory according to the rule of index $\geq .70$.

The item-total correlations for the French Eyes Test-31 present values in the small range for effect size (Cohen, 1988). The mean corrected item-total polyserial correlations are low, as are the mean inter-item tetrachoric correlations. These results are related to the method chosen to select reliable items, a process that was named the *consensus method* in the original article by Baron-Cohen et al. (2001). By definition, the percentage of correct responses to an item cannot be inferior to 50%, inducing necessarily a small variance. Therefore, selected items with less than fifty percent of correct responses but with good psychometric properties is justified, with the aim to increase the difficulty of the test. Indeed, the variance for a binary variable is mathematically linked to its mean and therefore the range of item-total Pearson correlations for binary items is frequently restricted if the mean of the items is not .50 (McDonald, 1999). The consensus method requires that the variance of each item is small, because all the items should have a mean of correct response superior to .50. This bias (low variance of items) is clearly apparent in the change of the size of omega when it is computed as an ordinal coefficient: .79 versus .61 for the "classic" omega coefficient.

This surprising change could be explained by the dichotomous nature of the items and their low variance. Indeed, the tetrachoric correlations appear to be a good measure for the low covariance and allow the omega coefficient to be probably closer to its true value, contrary to the Pearson correlations. A proof of the interpretation of the high guessing as caused by the design of the items and the small variance induced, is that one of the two items with the smallest guessing parameters is one with an associated high level of choice of the distractors (31.7%); according to Baron-Cohen et al. (2001), this item should be removed. However, the results suggest that this item should not be excluded.

Confirmatory Factor Analysis and parallel analyses, two among the strongest methods to confirm dimensionality, support the unidimensionality of the French Eyes-Test-31. Our data support the hypothesis that a unique latent variable explains the results on the test and this unique ability is probably affective ToM. This hypothesis is further confirmed by the positive relationship between the French Eyes Test-31 and the FERT. Item 4 and 35 present a covariance that cannot be explained uniquely by the general factor, probably because of the use of the same target word (“charmeur”).

Responses to the French Eyes-Test-31involve various abilities, like the understanding of the word, visual face perception, the ability to detect the target word, elimination of the implausible words or distractors (Johnston et al., 2008), and probably the *g factor*. Peterson and Miller (2012) found that 25 % of the variance in the Eyes Test was explained by an index of the *Verbal Intelligence Quotient* (VIQ). In a meta-analysis, Baker et al. (2014) showed that this result is strong and involves both verbal and performance intelligence quotients. However, their correlation between intelligence and the score at the Eyes Test was not high with r = .24, which is a relatively small effect size.

The processes involved in the success on the Eyes Test are complicated by the high level of guessing revealed in this study. Indeed, for the first time, the high level of guessing in

the Eyes-Test has been studied. Our results confirm those of Johnston et al. (2008), who experimentally studied this effect. Therefore, we confirm the importance of guessing in this task. The high level of both guessing and correct target recognition affects probably negatively the sensitivty of the test. The Bayesian item response model analysis showed its potential to inform us about the characteristics of the test. The consensus method for the choice of items probably induces a high level of successful guessing.

4.2.Convergent and discriminant validities

A result of our study is that FERT was significantly positively correlated with verbal intelligence, as well as with the French Eyes Test-31 ($p<.05$). This result is surprising, because the words used in the French Eyes Test-31 seem far more complex than the words used for basic emotions, which are universal; therefore one could presume that verbal intelligence would be involved in the French Eyes Test-31, by the simple need to understand the meaning of complex words. However, this does not appear to be the case. This result suggests that it is not only verbal comprehension that explains the association between the Eyes Test and intelligence in previous studies, but rather a more general level of intelligence, like the *g factor* (e.g. Sternberg & Grigorenko, 2002). The present study supports the hypothesis of an implication of a g-like factor, explaining its involvement even in the simpler facial emotion recognition test.

4.3.Test-retest reliability

The ICC coefficient confirms the test-retest reliability, with a value superior to .70, which is a sufficiently high level of test-retest reliability. Therefore, the reliability of the Eyes Test-31 is supported by the test-retest procedure.

CONCLUSION

The French Eyes-Test-31 is a reliable and valid test, with discriminant items and good psychometric properties. Its reliability coefficients (ordinal ω) are within the range of acceptable values (i.e. ≥ .70); its unidimensionality is supported by the results. The present results support the use of “ordinal” reliability coefficients. A unique latent ability seems to be involved in the success on this test, the affective ToM. Evidence is provided for the convergent and discriminant validities of the French Eyes Test-31, as well as its test-retest reliability. Its optimal properties are reflected in the quasi absence of "ceiling effects" (only three of the normal participants out of 661 attained the maximum score). This conclusion is supported by the computed significant discriminant coefficient (3-PL model) of all the selected items, which all attest that they participate to the test’s ability to discriminate between high and low functioning individuals. However, the test could be improved, given the high level of correct guessing associated with this test. In conclusion, the French Eyes Test-31 is ‘an advanced test of theory of mind’. It can be used in the general population to search, for example, for the endophenotypes of neurological or psychiatric syndromes in the relatives of patients, in contrast to many other tests of ToM, which show strong ceiling effects.

REFERENCES

Baker, C. A., Peterson, E., Pulos, S., & Kirkland, R. A. (2014). Eyes and IQ: A meta-analysis of the relationship between intelligence and "Reading the Mind in the Eyes". *Intelligence*, *44*, 78–92.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome. *Journal of Child Psychology and Psychiatry*, *38*(7), 813–822. https://doi.org/10.1111/j.1469-7610.1997.tb01599.x

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

Beauducel, A., & Herzberg, P. Y. (2009). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA, (October 2014), 37–41. https://doi.org/10.1207/s15328007sem1302

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Cho, S.-J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement*, *69*(5), 748–759.

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software*, *36*(1), 1–34.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied psychology*, *68*(3), 363.

Ekman, P. (1992a). An argument for basic emotions. *Cognition & Emotion*, *6*(3‑4), 169–200.

Ekman, P. (1992b). Are there basic emotions? *Psychological Review*.

Ekman, P., & Friesen, W. V. (1975). *Pictures of facial affect*. Consulting Psychologists Press.

Fox, J. (2010). *Bayesian item response theory*. New York: Springer.

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, *17*(3), 1–13.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments* (Vol. 196). Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Hallerbäck, M. U., Lugnegard, T., Hjärthag, F., & Gillberg, C. (2009). The Reading the Mind in the Eyes Test: test–retest reliability of a Swedish version. *Cognitive neuropsychiatry*, *14*(2), 127–143.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424‑453. https://doi.org/10.1037/1082-989X.3.4.424

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1‑55. https://doi.org/10.1080/10705519909540118

Johnston, L., Miles, L., & McKinlay, A. (2008). A critical review of the eyes test as a measure of social-cognitive impairment. *Australian Journal of Psychology*, *60*(3), 135‑141. https://doi.org/10.1080/00049530701449521

Kirkland, R. A., Peterson, E., Baker, C. A., Miller, S., & Pulos, S. (2013). Meta-analysis Reveals Adult Female Superiority in" Reading the Mind in the Eyes Test". *North American Journal of Psychology*, *15*(1).

Kruschke, J. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Elsevier Science. Consulté à l'adresse http://books.google.co.uk/books?id=FzvLAwAAQBAJ

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores.*

Lorenzo-Seva, U., & Ferrando, P. J. (2013). FACTOR 9.2 A Comprehensive Program for Fitting Exploratory and Semiconfirmatory Factor Analysis and IRT Models. *Applied Psychological Measurement*, *37*(6), 497–498.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049‑3067. https://doi.org/10.1002/sim.3680

McDonald, R. P. (1999). *Test theory: A unified treatment.* Psychology Press.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64‑82. https://doi.org/10.1037/1082-989X.7.1.64

Peterson, E., & Miller, S. F. (2012). The Eyes Test as a Measure of Individual Differences: How much of the Variance Reflects Verbal IQ? *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00220

Pinkham, A. E., Penn, D. L., Green, M. F., Buck, B., Healey, K., & Harvey, P. D. (2014). The Social Cognition Psychometric Evaluation Study: Results of the Expert Survey and RAND Panel. *Schizophrenia Bulletin*, *40*(4), 813‑823. https://doi.org/10.1093/schbul/sbt081

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125). Technische Universit at Wien.

Preti, A., Vellante, M., & Petretto, D. R. (2017). The psychometric properties of the "Reading the Mind in the Eyes" Test: an item response theory (IRT) analysis. *Cognitive Neuropsychiatry*, *22*(3), 233‑253. https://doi.org/10.1080/13546805.2017.1300091

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Consulté à l'adresse http://www.R-project.org/

Raven, J. C., & Deltour, J. (1998). *Echelle de vocabulaire Mill Hill*. Ed. et applications psychologiques.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2). Consulté à l'adresse http://www.lce.esalq.usp.br/arquivos/aulas/2013/encontro_ppg/Lucia/paper.pdf

Sanvicente-Vieira, B., Kluwe-Schiavon, B., Wearick-Silva, L. E., Piccoli, G. L., Scherer, L., Tonelli, H. A., & Grassi-Oliveira, R. (2013). Revised Reading the Mind in the Eyes Test (RMET)-Brazilian version. *Revista Brasileira de Psiquiatria*, (AHEAD), 000–000.

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia*, *45*(13), 3054‑3067. https://doi.org/http://dx.doi.org/10.1016/j.neuropsychologia.2007.05.021

Shamay-Tsoory, S. G., Shur, S., Barcai-Goodman, L., Medlovich, S., Harari, H., & Levkovitz, Y. (2007). Dissociation of cognitive from affective components of theory of mind in schizophrenia. *Psychiatry research*, *149*(1), 11–23.

Sternberg, R. J., & Grigorenko, E. L. (2002). *The general factor of intelligence: How general is it?* Psychology Press.

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., & Preti, A. (2013). The “Reading the Mind in the Eyes” test: systematic review of psychometric properties and a validation study in Italy. *Cognitive neuropsychiatry*, *18*(4), 326–354.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research*, *19*(1), 231–240.

Yıldırım, E. A., Kaşar, M., Güdük, M., Ateş, E., Küçükparlak, I., & Ozalmete, E. O. (2011). Investigation of the reliability of the « reading the mind in the eyes test » in a Turkish population. *Turkish journal of psychiatry*, *22*(3), 177‑186.

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. University of California Los Angeles.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment*, *31*(4), 231‑237. https://doi.org/10.1027/1015-5759/a000309

Table 1: 3-PL Bayesian analysis on 35 items, medians and 95 % credible interval of alpha coefficients

English denomination	French number	2.50%	50%	97.50%
practice	Item 1	0.163	0.601	1.775
Item 1	Item 2	1.033	1.689	2.888
Item 2	Item 3	0.282	0.762	2.126
Item 3	Item 4	0.771	1.264	2.33
Item 4	Item 5	0.187	0.845	2.639
Item 5	Item 6	0.268	0.68	1.992
Item 6	Item 7	0.612	1.042	1.971
Item 7	Item 8	0.143	0.453	2.174
Item 8	Item 9	0.234	0.548	1.787
Item 9	Item 10	0.558	0.948	1.704
Item 11	**Item 11**	**-0.181**	**0.234**	**1.957**
Item 12	Item 12	0.326	0.656	1.317
Item 14	Item 13	0.343	0.862	2.276
Item 15	Item 14	0.589	0.949	1.648
Item 16	Item 15	0.396	0.911	2.351
Item 18	Item 16	0.205	0.78	1.621
Item 20	Item 17	0.516	1.299	2.909
Item 21	Item 18	1.185	1.98	3.269
Item 22	Item 19	0.4	1.025	2.452
Item 23	**Item 20**	**-0.129**	**0.16**	**1.515**
Item 24	Item 21	0.082	0.34	1.168
Item 25	Item 22	0.332	0.627	1.268
Item 26	Item 23	0.689	1.063	1.748
Item 27	**Item 24**	**-0.29**	**0.096**	**2.249**
Item 28	Item 25	0.233	0.724	2.274
Item 29	Item 26	0.405	1.069	2.595
Item 30	Item 27	0.729	1.155	2.082
Item 31	Item 28	0.121	0.49	1.942
Item 32	Item 29	0.134	0.425	1.317
Item 34	Item 30	0.026	0.471	2.254
Item 36	Item 31	0.073	0.568	1.582
Item 13	Item 32	0.425	0.937	2.104
Item 33	Item 33	0.415	0.824	1.869
Item 36	**Item 34**	**-0.051**	**0.211**	**2.246**
child version#	Item 35	0.823	1.296	2.197

In bold, the 4 items with credible intervals including 0 and excluded from the final version of the French Eyes-Test-31; with the correspondence with English item numbers

Table 2. Percent of responses for targets and distractors in the Validation of the French Eyes Test-31

English denomination	French number	Target	Foil 1	Foil 2	Foil 3
practice	1	81.9	7.5	6.2	4.4
Item 1	2	93.1	2.9	2.7	1.3
Item 2	3	84.0	8.7	3.7	3.7
Item 3	4	77.7	17.7	3.3	1.3
Item 4	5	82.9	8.7	5.8	2.7
Item 5	6	76.9	12.1	10.4	0.6
Item 6	7	67.9	13.3	11.9	6.9
Item 7	**8**	**47.5**	**26.3**	**23.7**	**2.5**
Item 8	9	59.0	**31.7**	6.9	2.3
Item 9	10	87.1	7.1	5.2	0.6
Item 12	12	77.3	17.1	4.0	1.5
Item 14	13	88.5	8.7	1.5	1.3
Item 15	14	81.3	9.8	4.8	4.0
Item 16	15	80.6	15.0	2.5	1.9
Item 18	16	91.3	4.0	3.1	1.5
Item 20	17	92.3	4.0	3.7	0.0
Item 21	18	97.3	1.7	1.0	0.0
Item 22	19	91.0	4.4	3.8	0.8
Item 24	21	56.7	16.2	15.2	11.9
Item 25	22	63.5	18.1	10.2	8.3
Item 26	23	78.3	12.5	6.0	3.3
Item 28	25	81.9	12.1	3.7	2.3
Item 29	**26**	**44.8**	**33.1**	**16.5**	**5.6**
Item 30	27	80.4	7.3	6.7	5.6
Item 31	28	66.7	16.2	10.2	6.9
Item 32	29	66.5	23.7	7.5	2.3
Item 34	30	78.8	12.1	5.0	4.0
Item 36	31	88.7	8.5	2.9	0.0
Item 13	32	89.2	6.2	4.0	0.6
Item 33	33	70.8	22.1	4.0	3.1
child version#	35	76.5	10.0	8.5	5.0

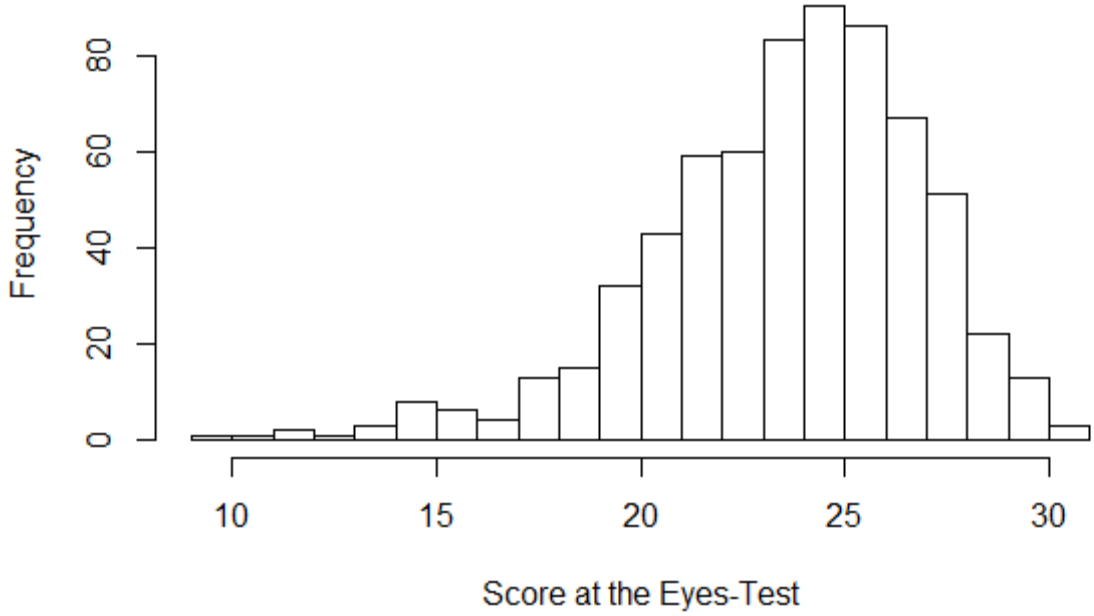
* In boldface, items with less than 50 % of correct responses or with higher than 25 % of distractor choice

item adapted from the child version of the English Eyes Test

Table 3. Results of the Bayesian Item Response Model of the French Eyes Test-31, quantiles of the distribution of guessing parameters

Item	2.50%	25%	50%	75%	97.50%
Item 1	**0.373**	0.591	0.657	0.713	0.802
Item 2	0.107	0.48	0.634	0.733	0.843
Item 3	**0.325**	0.577	0.657	0.72	0.803
Item 4	0.042	0.261	0.395	0.498	0.634
Item 5	**0.495**	0.666	0.727	0.777	0.83
Item 6	0.212	0.459	0.55	0.628	0.726
Item 7	0.035	0.214	0.328	0.421	0.548
Item 8	0.005	0.063	0.154	0.315	0.449
Item 9	0.026	0.17	0.278	0.392	0.546
Item 10	0.074	0.377	0.526	0.635	0.765
Item 12	0.103	0.366	0.478	0.567	0.69
Item 13	**0.349**	0.617	0.698	0.756	0.825
Item 14	0.042	0.271	0.415	0.532	0.677
Item 15	**0.278**	0.54	0.631	0.698	0.775
Item 16	**0.454**	0.721	0.792	0.838	0.892
Item 17	**0.564**	0.765	0.82	0.857	0.901
Item 18	0.106	0.514	0.687	0.794	0.899
Item 19	**0.482**	0.73	0.795	0.839	0.889
Item 21	0.011	0.087	0.157	0.253	0.473
Item 22	0.031	0.189	0.299	0.4	0.553
Item 23	0.017	0.15	0.268	0.383	0.547
Item 25	**0.32**	0.534	0.612	0.68	0.757
Item 26	0.029	0.199	0.287	0.341	0.408
Item 27	0.047	0.286	0.429	0.542	0.685
Item 28	0.101	0.292	0.386	0.493	0.626
Item 29	0.032	0.173	0.264	0.362	0.54
Item 30	**0.344**	0.51	0.576	0.662	0.752
Item 31	**0.519**	0.706	0.756	0.795	0.861
Item 32	**0.324**	0.634	0.725	0.784	0.854
Item 33	0.048	0.257	0.378	0.475	0.595
Item 35	0.018	0.151	0.263	0.368	0.522

Figure 1. Distribution of total scores on the Validation of the French Eyes-Test 31 (N = 661)



Appendix 1: Words used in the Validation of the French Eyes Test-31

English denomination	French number	Choice 1	Choice 2	Choice 3	Choice 4
practice	1	jaloux	**paniqué**	arrogant	haineux
Item 1	2	**malicieux**	consolateur	irrité	ennuyé
Item 2	3	terrifié	**bouleversé**	arrogant	agacé
Item 3	4	plaisantant	troublé	**charmeur**	convaincu
Item 4	5	plaisantant	**insistant**	amusé	détendu
Item 5	6	agacé	sarcastique	**soucieux**	amical
Item 6	7	consterné	**rêveur**	impatient	inquiet
Item 7	8	se justifiant	amical	**mal à l'aise**	déprimé
Item 8	9	**découragé**	soulagé	timide	impatient
Item 9	10	agacé	hostile	horrifié	**préoccupé**
Item 12	12	indifférent	embarrassé	**sceptique**	déprimé
Item 14	13	irrité	déçu	abattu	**accusateur**
Item 15	14	**contemplatif**	confus	encourageant	amusé
Item 16	15	irrité	**pensif**	encourageant	compatissant
Item 18	16	**décidé**	amusé	consterné	ennuyé
Item 20	17	impatient	**amical**	coupable	horrifié
Item 21	18	embarrassé	**charmeur***	confus	paniqué
Item 22	19	**préoccupé**	reconnaissant	insistant	implorant
Item 24	21	**méditatif**	irrité	impatient	hostile
Item 25	22	paniqué	incrédule	découragé	**intéressé**
Item 26	23	alarmé	timide	**hostile**	anxieux
Item 28	25	**intéressé**	plaisantant	affectueux	satisfait
Item 29	26	impatient	consterné	irrité	**méditatif**
Item 30	27	reconnaissant	**flirtant**	hostile	déçu
Item 31	28	honteux	**confiant**	plaisantant	abattu
Item 32	29	**grave**	honteux	perplexe	alarmé
Item 34	30	consterné	déconcerté	**méfiant**	terrifié
Item 36	31	honteux	nerveux	**soupçonneux**	indécis
Item 13	32	décidé	**espérant**	menaçant	arrogant*
Item 33	33	satisfait	coupable	rêveur	**préoccupé**
child version#	35	haineux*	nerveux*	plaisantant*	**charmeur***

*: word meaning changed from the original English version (Baron-Cohen et al., 2001) for psychometric reasons.
#: item from the English child version of the Eyes Test with new words (targets and distractors)
Words in boldface are target words.