PSIHOLOGIJA, 2017, Vol. 50(4), 483–502
DOI: https://doi.org/10.2298/PSI170504010D

UDC 159.923.072.59

Psychometric evaluation and validation of the Serbian version of "Reading the Mind in the Eyes" test

Jelena Đorđević[1], Marko Živanović[2], Aleksandra Pavlović[3], Goran Mihajlović[4], Ivana Stašević Karličić[1], and Dragan Pavlović[2,5]

[1]*Clinic for Psychiatric Disorders, "Dr Laza Lazarević", Belgrade, Serbia*
[2]*Department of Psychology, Faculty of Philosophy, University of Belgrade, Serbia*
[3]*Neurology Clinic, Clinical Center of Serbia, Belgrade, Serbia*
[4]*Faculty of Medical Sciences, University of Kragujevac, Serbia*
[5]*Faculty for Special Education and Rehabilitation, University of Belgrade, Serbia*

"Reading the Mind in the Eyes" test (*RMET*) is one of the most popular and widely used measures of individual differences in Theory of Mind (ToM) capabilities. Despite demonstrating good validity in differentiating various clinical groups exhibiting ToM deficits from unimpaired controls, previous studies raised the question of the *RMET's* homogeneity, latent structure, and reliability. The aim of this study is to provide evidence on psychometric properties, latent structure, and validity of the newly adapted Serbian version of the *RMET*. In total, 260 participants (61.9% females) took part in the study. The sample consisted of both unimpaired controls (76.5%), and a clinical group of participants that are believed to demonstrate ToM deficits (23.5%), namely, persons diagnosed with schizophrenia and bipolar disorder (54.1% females). *RMET* has demonstrated fair psychometric properties (*KMO* = .723; α = .747; *H1* = .076; *H5* = .465), successfully differentiating between clinical group and control [F (1,254) = 26.175, p <.001, η^2_p = .093], while typical gender differences in performance were found only in control group. Tests of several models based on the previous literature revealed that the affect-specific factors underlying performance on *RMET* demonstrate poor fit. The best fitting model obtained included reduced scale with a single-factor underlying the test's performance (*TLI* = .953, *CFI* = .958, *RMSEA* = .020). Based on the fit parameters we propose 18-item short-form of the Serbian version of *RMET* (*KMO* = .797; α = .728; *H1* = .129; *H5* = .677) for economic, reliable and valid measurement of ToM abilities.

Keywords: Reading the Mind in the Eyes, *RMET*, Theory of Mind, ToM, psychometric evaluation

Corresponding author: jelenadjordjevic2000@yahoo.com

Acknowledgement. Second author was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (179018)

Highlights:

- The aim of the study was to provide evidence on psychometric properties, latent structure, and the validity of the Serbian version of *RMET.*
- *RMET* demonstrated fair psychometric properties and satisfactory validity in differentiating between the clinical group and the healthy controls.
- Several models tested revealed that the affect-specific factors underlying performance on *RMET* demonstrated poor fit, while the reduced single-factor models exhibited much better fit.
- The short-form of the Serbian version of *RMET* measuring single-factor of general ToM abilities is proposed.

Social cognition is a mental operation which lies at the basis of altruistic behavior, caused by empathizing or understanding hints made by other people which show a need for concealment, sharing and help (Mussen & Eisenberg, 1977). According to Addington there are four domains of social cognition: Theory of Mind (ToM), attributive style, perception of emotions, and social observation (Addington, Penn, Woods, Addington, & Perkins, 2008). Social cognition can be divided into lower-level processes such as recognition and perception of socio-emotional signs including facial expressions, depth of voice, gestures; and higher-level processes such as inferring conclusions about mental states of others (that is ascribing mental states), empathy and emotional regulation (Ochsner, 2008). The capacity for emotional investment in relationships and moral standards indicates the orientation of the society focused on the need, as opposed to investing in values, ideals, and interpersonal relations. Damage to social cognition is observed in different clinical entities – from pervasive disorders to endogenous psychosis, eating and personality disorders. ToM tests are frequently used for assessment of social cognition.

Theory of mind

Theory of mind (ToM) is a concept that describes people's ability to understand and describe the mental states of other people, their intentions and beliefs (Premack & Woodruff, 1978). More specifically, ToM studies the psychological processes that serve to understand others or make mental boundaries between self and others (Doherty, 2009). Scholars suggest that the basis of ToM is a kind of mental modeling in which the simulator uses his mental frame of mind as an analog model simulating the object (Gordon, 1986).

ToM is called a *theory* because it assumes that mental states of others are not directly detectable but must be generated through predictions about how others think and will behave. This theory was originally developed to describe the behavior of chimpanzees (Premck & Woodruff, 1978), and then was expanded to describe the development of children and their ability to predict the

perspective of others (Wellman, Cross, & Watson 2001). Later on this model was applied in description of the social and communicative deficits in specific clinical populations, mostly from the spectrum of autism (Baron-Cohen, Leslie, & Frith, 1985). It is considered conceptually similar or equivalent to cognitive empathy (Baron-Cohen et al., 2015) because both constructs include conclusions about the mental state of another person. There are two disciplines studying ToM: social science, exploring the neural basis of ToM and developmental psychology, interested in how these capabilities develop (Mahy, Moses, & Pfeifer, 2014). There are four major theories of ToM development in children: modularity, simulation, executive and theory theories (Mahy et al., 2014).

Neuroimaging studies provided some evidence on the neural basis of ToM. Functional magnetic resonance imaging studies assess the neural substrates of ToM in situations where respondents are thinking about their own or someone else's mental state. These studies demonstrated the activation of the posterior superior temporal sulcus and temporoparietal junction, medial prefrontal cortex, temporal poles and precuneus in ToM type tasks (Frith, 2007). Affective ToM seems to be based on a phylogenetically older emotional system in the lower frontal gyrus, while the cognitive ToM is likely dependent on the functioning of the ventromedial prefrontal gyrus (Shamay-Tsoory, Harari, Aharon-Peretz, & Levkovitz, 2010). The role of the ventromedial prefrontal cortex is controversial given the numerous connections of ventromedial prefrontal cortex with other regions such as the amygdala, superior temporal sulcus and anterior insula (Shamay-Tsoory, Tibi-Elhanany, & Aharon-Peretz 2006).

"The blindness of the mind" is the opposite of ToM. That is a cognitive disorder characterized by an inability to ascribe a mental state to self or another person. This feature appears in people with Asperger's syndrome, autism, and schizophrenia as well as in other disorders that show a deficit of social insight. A person with this disorder is unable to understand or predict mental states of other people (Frith, 2001, Pijnenborg, Spikman, Jeronimus, & Aleman, 2013).

While the ToM is usually considered as one unitary construct, some authors have described it as multiple constructs which include perception, attention, beliefs, desires, intentions, and emotions (Astington, 2003). According to this approach, the tests used for assessment of ToM should be multiple, assessing subconstructs (Slaughter & Repacholi, 2003). However, in practice, researchers and clinicians use unidimensional tests such as the "Reading the Mind in the Eyes" test.

Reading the Mind in the Eyes test

The "Reading the Mind in the Eyes" test (*RMET*) is considered to be a measure of nonverbal aspects of ToM. *RMET* is commonly used for ToM assessment both in general and clinical populations, with a special focus on the autistic spectrum disorders. The test is designed to measure the first level of ToM – attribution, which identifies the relevant mental state, as opposed to the second level in which the content of mental state is inferred (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). This test has been developed under

the assumption that ToM heavily relies on the perception of eye gaze of the person being observed (Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997; Baron-Cohen et al., 2001) since it's considered as an important aspect of social interaction and communication (Emery, 2000).

The original version of the test (Baron-Cohen et al., 1997) consists of a series of 25 photographs depicting the area around the eyes with two descriptors of mental states presented with each photography. The participant's task is to select an alternative s/he considers to be the most suitable description of feelings or thoughts expressed by a person on a photograph. In order to resolve some of the issues the test was facing (see Baron-Cohen et al., 2001), revised version of the instrument was designed (Baron-Cohen et al., 2001). The second version of *RMET* consists of 36 male and female photographs (approximately equalized) of the area around the eyes with four descriptors of mental states offered, out of which only one is the correct description of feelings or thoughts expressed by a person on the photo. So far, this test has been adapted and translated into variety of languages, e.g. Italian (Vellante et al., 2013), French (Prevost et al., 2014), Romanian (Miu, Pana, & Avram, 2012), Bosnian (Schmidt & Zachariae, 2009), Spanish (Fernández-Abascal, Cabello, Fernández-Berrocal, & Baron-Cohen, 2013), German (Pfaltz et al., 2013), Turkish (Girli, 2014; Yildirim et al., 2011), Swedish (Hallerbäck, Lugnegård, Hjärthag, & Gillberg, 2009), Japanese (Adams et al., 2009; Kunihira, Senju, Dairoku, Wakabayashi, & Hasegawa, 2006), Persian (Khorashad et al., 2015), etc.

Despite demonstrating good validity in differentiating between people with autistic spectrum disorders (e.g. Baron-Cohen et al., 1985, 1997, 2001, 2015; Lai et al., 2012; Lombardo, Barnes, Wheelwright, & Baron-Cohen, 2007; Losh et al., 2009), various clinical groups (e.g. Bora, Bartholomeusz, & Pantelis, 2016; Bora, Yucel, & Pantelis, 2009; Hawken et al., 2016; Maurage et al., 2009; Schmidt & Zachariae, 2009), and unimpaired controls, as well as the evidence on its convergent validity (Baron-Cohen et al., 2001; Ferguson & Austin, 2010; Olderbak et al., 2015; Torralva, Roca, Gleichgerrcht, Bekinschtein, & Manes, 2009; Vellante et al., 2013; Voracek & Dressler, 2006) previous studies raised the question of *RMET's* homogeneity and reliability.

Previous studies found that the test shows low internal consistency, approximately falling within the range of .40 – .70 (see Harkness, Jacobson, Duong, & Sabbagh, 2010; Khorashad et al., 2015; Olderbak et al., 2015; Prevost et al., 2014; Ragsdale & Foley, 2011; Vellante et al., 2013; Voracek & Dressler, 2006). On the other hand, the test demonstrates acceptable test-retest reliability (see Hallerbäck et al., 2009; Khorashad et al., 2015; Prevost et al., 2014; Vellante et al., 2013; Yildirim et al., 2011). Additionally, there is a limited evidence of *RMET's* homogeneity and its factor structure. Namely, it is unclear whether this test is unidimensional – measuring a unitary construct of ToM as authors suggested (Baron-Cohen et al., 2001), or different affect-specific factors (see Harkness et al., 2005; Konrath, Corneille, Bushman, & Luminet 2014; Olderbak et al., 2015). In order to resolve some of the aforementioned issues, previous studies proposed several *RMET* reduced scales (e.g. Konrath et al., 2014; Olderbak et al., 2015).

Present study

This study aims to explore aforementioned issues through an examination of psychometric properties, latent structure, and validity of the Serbian adaptation of *RMET*. On the following pages, we present a psychometric evaluation of newly adapted Serbian version of the *RMET* and provide a comparison of its psychometric quality with other adaptations made. This study addresses notions on latent structure of *RMET* facing several concurrent models found in previous literature, trying to establish whether the object of *RMET's* measurement is unidimensional and general in nature, or multidimensional and affect-specific. Moreover, validity of the instrument was examined through testing its predictive power in differentiating between entities that are supposed to demonstrate ToM specific deficits, namely persons suffering from schizophrenia and bipolar disorder (Bora et al., 2016; Bora et al., 2009) and unimpaired controls, as well as through testing typically observed gender differences, i.e. "female superiority" in the performance on *RMET* (see Baron-Cohen et al., 2015; Baron-Cohen, Knickmeyer, & Belmonte, 2005; Baron-Cohen et al., 2001; Khorashad et al., 2015; Schiffer, Pawliczek, Muller, Gizewski, & Walter, 2013; Vellante et al., 2013). Finally, based on the results obtained, we propose a short, economic version of the instrument and contrast it with other short versions of the test suggested in the previous literature.

Method

Participants

A sample of 260 participants, age range 18 to 64 (M = 32.44, SD = 11.47; 61.9% females) took part in the study. Participants' years of education varied from 8 to 22, with the mean value of approximately 14 years (M = 13.66, SD = 2.59). In order to cover full spectrum of the variability of the construct measured, and to test the diagnostic validity of the instrument, the sample consisted of participants from both the student and the general population (76.5%), as well as the clinical population (23.5%; 54.1% females). More specifically, persons diagnosed with schizophrenia (49.2%) and bipolar disorder (50.8%) were included in the sample since previous studies showed that these entities demonstrate ToM-specific deficits (Bora et al., 2016; Bora et al., 2009). Subjects participated in the study on a voluntary basis and have signed an informed consent.

Instrument

Translation and cross-cultural adaptation of the test followed the instructions of the Autism Research Centre (ARC; www.autismresearchcentre.com) and relied on the experience of other researchers who have had the same adaptation done in other cultural environments. Adaptation of the original instrument (Baron-Cohen et al., 2001) was carried out using standard backward translation method, i.e. by researchers bilingual in English and Serbian, as well as by the professional translator. Preliminary Serbian version was tested on 40 subjects, after which, with minimal corrections, the test was submitted to the ARC for approval. Upon the approval, the test was administrated to participants in line with the instructions provided by the ARC (Baron-Cohen et al., 2001).

The revised version of the "Reading the Mind in the Eyes" test (*RMET*) (Baron-Cohen et al., 2001) consists of 36 photographs which present eyes region of different individuals (19 male stimuli, 17 female stimuli). Each of the photographs is presented along with the

four descriptors of complex mental states (Figure 1). Participants' task was to, among the descriptors offered, select the one which seems to be the most appropriate description of feelings or thoughts expressed by the individuals presented in the photograph. Among the descriptors offered, within each item, there is one target word and three foils.

Figure 1. Example of the item from *RMET*

Procedure

Following the practice section in which participants were familiarized with the task, they were successively presented with 36 eyes photographs each followed by four descriptors offered. Participants' task was to select the most appropriate one among four descriptors of mental state (feelings or thoughts) of a person presented in the photo. Glossary of mental states has been provided and participants could consult it at any time during testing. Testing was not time-limited, but participants were given an instruction not to contemplate too much on individual items.

Results

Table 1 displays percentage of participants who have chosen each option within every item. As shown, the proportion of participants who have chosen target words ranges from .46 to .91, with items in most cases being successfully solved by at least 50% of participants. Furthermore, it can be noted that some items exhibited specific patterns of option selection. More specifically, it is evident that most of the items have one salient distractor that competes with the target word while other options are seldom chosen. For example, the odds of option 4 being (wrongly) selected as a target word within item 3 is 9 times higher than for option 1 and 14 times higher than for option 2. Similar disproportion can be found within item 6, for example. The number of items with more than one dominant option competing for the correct answer is disproportionally low (for example, items 8, 9, 11, 13, 15, etc.).

Twenty-eight out of 36 items have shown to fall within the range of item difficulties obtained in previous studies (Table 1). Five of those items have shown to be easier, while three items proved to be more difficult compared to the other versions of *RMET*. However, in six out of eight items, aforementioned deviations have not exceeded 5% of the increase/decrease in items' difficulty as compared to other versions of the instrument.

Table 1
Percentage of participants who have chosen each option in each item (item difficulty/target words are marked bold), item difficulties in previous studies, stimulus gender, and emotional valence of stimuli

items	option 1	option 2	option 3	option 4	item difficulties in previous studies[1] (range)	stimulus gender	emotional valence[2]
1	**49.8**	19.7	23.3	7.2	70.0 (53.8 – 85.2)	M	positive
2	14.5	**65.5**	14.0	6.0	67.7 (49.4 – 83-8)	M	negative
3	3.2	2.0	**66.7**	28.1	77.6 (53.4 – 93.0)	F	neutral
4	2.0	**79.7**	6.5	11.8	70.7 (57.0 – 81.1)	M	neutral
5	4.3	12.6	**79.9**	3.2	73.5 (58.2 – 92.5)	M	negative
6	2.0	**72.2**	23.4	2.4	75.4 (69.0 – 80.3)	F	positive
7	8.9	29.4	**50.8**	10.9	47.5 (18.7 – 68.0)	M	neutral
8	**87.5**	5.9	2.3	4.3	74.7 (67.0 – 88.0)	M	neutral
9	5.2	6.7	6.4	**81.7**	78.3 (61.1 – 90.5)	F	neutral
10	**75.2**	13.2	9.2	2.4	65.4 (43.9 – 76.0)	M	neutral
11	3.6	11.5	**75.4**	9.5	67.8 (52.1 – 77.8)	M	negative
12	15.9	2.4	**79.7**	2.0	77.2 (63.2 – 87.7)	M	neutral
13	4.4	**80.7**	3.6	11.3	63.5 (34.0 – 80.8)	M	neutral
14	20.3	4.4	4.0	**71.3**	83.8 (73.4 – 93.6)	M	negative
15	**88.2**	3.5	3.5	4.8	81.4 (69.7 – 86.9)	F	neutral
16	2.0	**78.4**	2.8	16.8	76.8 (59.4 – 85.8)	M	positive
17	**77.7**	17.8	0.9	3.6	55.2 (48.0 – 65.6)	F	negative
18	**90.9**	3.9	2.4	2.8	83.4 (58.2 – 96.4)	F	neutral
19	7.6	17.7	7.6	**67.1**	52.8 (38.5 – 69.7)	F	neutral
20	13.5	**71.5**	13.4	1.6	85.0 (73.5 – 92.0)	M	positive
21	11.2	**78.9**	8.3	1.6	63.7 (39.4 – 86.0)	F	positive
22	**78.6**	2.4	8.7	10.3	81.1 (70.8 – 90.5)	F	negative
23	2.5	6.5	**59.2**	31.8	59.6 (37.0 – 77.9)	M	negative
24	**72.4**	9.1	7.1	11.4	68.7 (57.4 – 84.0)	M	neutral
25	4.0	22.1	19.3	**54.6**	61.1 (39.3 – 76.0)	F	positive
26	3.1	3.2	**68.9**	24.8	73.2 (65.6 – 78.1)	M	negative
27	1.6	**76.1**	10.6	11.7	59.8 (47.5 – 67.1)	F	negative
28	**74.5**	1.6	9.6	14.3	70.5 (47.0 – 89.7)	F	neutral
29	12.3	4.0	21.4	**62.3**	69.0 (38.2 – 84.6)	F	neutral
30	2.8	**89.8**	3.9	3.5	85.8 (76.9 – 91.0)	F	positive
31	4.0	**74.5**	8.0	13.5	58.0 (32.3 – 70.9)	F	positive
32	**82.7**	2.0	11.4	3.9	71.3 (50.0 – 80.0)	M	neutral
33	2.8	20.8	6.4	**70.0**	64.2 (54.0 – 77.4)	M	neutral
34	2.4	13.3	**79.0**	5.3	67.7 (54.7 – 77.0)	F	negative
35	38.7	**45.7**	9.8	5.8	56.7 (36.5 – 77.7)	F	negative
36	1.6	12.1	**70.9**	15.4	75.9 (65.8 – 87.5)	M	negative
mean		.70			.70 (.63 – .76)		

9 Item difficulties were calculated as mean percentiges of correct responses provided for German (Pfaltz et al., 2013), Turkish (Yildirim et al., 2011), Spanish (Fernández-Abascal et al., 2013), Italian (Vellante et al., 2013), French (Prevost et al., 2014), and Persian (Khorashad et al., 2015) adaptations of *RMET*, as well as values provided in the original publication (Baron-Cohen et al., 2001).

2 Classification of emotional valence of the target stimuli based on Harkness et al. (2005).

Analysis of variance revealed significant differences in item difficulties between different versions of the instrument [$F_{(7,245)} = 4.910$, $p < .001$, $\eta^2_p = .123$]. However, post hoc tests (Bonferroni) revealed that the items of the Serbian version did not show deviations from the original English version ($M_{diff} = 0.186$, $p = 1.00$), nor Turkish ($M_{diff} = 0.935$, $p = 1.00$), Spanish ($M_{diff} = -2.511$, $p = 1.00$), Italian ($M_{diff} = 3.958$, $p = .657$), French ($M_{diff} = 4.222$, $p = 1.00$), or German adaptations of the instrument ($M_{diff} = 5.008$, $p = 1.00$) In other words, the only significant deviation of the Serbian version of the instrument from any other was the one from the Persian adaptation of the *RMET* ($M_{diff} = 9.608$, $p < .01$).

Following the score calculation, descriptive statistic measures were obtained. The distribution of participants' scores has shown to be severely skewed ($zSK = -6.642$, $p < .01$), and elongated ($zKu = 4.439$, $p < .01$), indicating distortion of the distribution of scores from the normal toward higher scores in a leptokurtic manner ($K\text{-}S = 1.821$, $p < .01$). Individual scores on *RMET* were ranging from .14 to .97, with participants, on average succeeding to correctly solve .70 of the items ($SD = .14$).

In order to examine whether the Serbian version of *RMET* successfully discriminates between entities that are supposed to have ToM deficits and participants without those deficits, and to test[11] whether females perform better than males, two-factor analysis of covariance (ANCOVA) was performed, with age and number of years of education taken as covariates. Levene's test indicated equality of error variances across groups [$F_{(3,256)} = 0.358$, $p = .784$]. Results of ANCOVA indicated the significant main effect of group [$F_{(1,254)} = 26.175$, $p < .001$, $\eta^2_p = .093$], with clinical group performing significantly worse ($M = .58$, $SD = .16$) than unimpaired controls ($M = .74$, $SD = .10$). On the other hand, the main effect of gender [$F_{(1,254)} = 1.152$, $p = .284$], and group *x* gender interaction have not reached statistical significance [$F_{1,254)} = 1.777$, $p = .184$].

In order to cover full spectrum of the variability of the construct measured, the psychometric analysis was performed on both groups taken together. Psychometric characteristics of the test were calculated using the Rtt10g macro (Knežević & Momirović, 1996). Full-scale item sampling adequacy was .723 indicating lower representativeness of items sampled for measuring given ability. Internal consistency of the test has shown to be overall satisfying, $\alpha = .747$. Both average inter-item correlation ($H1 = .076$), as well as the proportion of variance accounted for by the first principal component relative to other components whose reliability is exceeding zero ($H5 = .465$) indicated lower test homogeneity.

Individual items' sampling adequacy has shown to vary between .240 and .884, with not a single item exceeding the level of .90 (Appendix A). The

11 Since distribution of scores demonstrated deviation from normal distribution this variable was normalized using Rankit formula (see Solomon & Sawilowsky, 2009).

proportion of variance of a given item predicted using the remaining of the test's items (item's reliability) has shown to be relatively low for most of the items, ranging from .083 to .332. On the other hand, both measures of item's internal validity have detected numerous items achieving moderate positive corrected item-total correlations (range .080 – .527), as well as a number of items whose correlations with the principal object of measurement can be considered satisfying (range .002 – .461). Yet, both measures indicated several items whose correlations with the object of measurement are achieving zero, pointing to their poor discriminative power and specificity in the context of remaining items.

In order to examine latent structure of the instrument, the exploratory factor analysis (EFA) was carried out. Maximum likelihood extraction was used along with Promax rotation of the axis. Guttman-Kaiser criterion suggested retention of 14 factors, while scree plot demarcated a slope change after the second factor. Following the latter criteria, the number of factors was fixed to two. Two retained factors accounted for 12.24% of the items' variance. Pattern matrix is presented in table 2. Correlation between two extracted factors has shown to be moderate (r = .453). Overviewing primary factor loadings, no interpretation by means of a type of mental state depicted in the image, or other stimuli characteristic seemed to be an appropriate explanation for the items' grouping.

Table 2
EFA's pattern matrix

	Factors			Factors	
	1	2		1	2
i1	.208	**-.372**	i19	.034	.328
i2	**.457**	-.118	i20	**.384**	-.035
i3	.018	**.115**	i21	**.360**	-.060
i4	.085	**.504**	i22	.222	**.278**
i5	.078	**.227**	i23	-.031	**.217**
i6	**.397**	-.135	i24	.105	**.166**
i7	.119	**.217**	i25	.192	**-.283**
i8	.108	**.323**	i26	-.074	**.247**
i9	**.355**	.002	i27	.155	**.263**
i10	.017	**.357**	i28	**.214**	.186
i11	-.082	**.508**	i29	**.240**	.083
i12	.180	**.268**	i30	**.229**	.138
i13	.337	.074	i31	**.559**	-.216
i14	**.291**	.080	i32	**.288**	-.057
i15	**.250**	.172	i33	**.141**	.075
i16	**.370**	.026	i34	**.276**	.116
i17	-.018	**.540**	i35	**.139**	.059
i18	**.360**	.101	i36	**.176**	.081

On the basis of theoretical expectations and previous empirical findings, several confirmatory factor analyses (CFA) were performed. Summary of the models tested is presented in Table 3 and factor loadings for seven models tested are presented in the Appendix B. First of all, through examination of the model fit for the single-factor full-scale solution we wanted to determine whether test is unidimensional, i.e. whether all the items successfully measure single latent trait as suggested by Baron-Cohen et al. (2001). Results have shown that the full-scale single-factor model has a poor fit, with the low average loading of .275 (Appendix B). Secondly, we tested the model obtained in the EFA with two interrelated factors underlying the performance on all the items. Estimated correlation between factors was high (r = .621), with average loadings of .321 and .279, for the first and second factor, respectively. Overall, this model has shown poor fit as well. Furthermore, four models, subsuming previous empirical findings were examined. Affect-specific three-factor model of positive, negative, and neutral factors (Harkenss et al., 2005) underlying performance on the *RMET* has shown poor fit, with average loadings of .254, .311, and .300, for positive,

neutral, and negative factor, respectively. Estimated correlations between factors have shown to be high for all the factor pairs – positive and negative (r = .666), positive and neutral (r = .763), and neutral and negative factor (r = .921). The two-factor model of positive and negative affect (Konrath et al., 2014) demonstrated somewhat better, but still unsatisfying fit, with very high positive estimated correlation between factors (r = .944), and average loadings for the first and the second factor of .218 and .336, respectively. On the other hand, reduced model of Konrath et al. (2014) has shown fair fit according to all fit indices, with the average loading of .292, while the model of Olderbak et al. (2015) demonstrated less good fit with the average loading of .302.

In order to get to the most appropriate and reliable model of the Serbian adaptation of *RMET*, which would be based on the theoretical expectation of a single factor underlying the ability measured we eliminated items which exhibited low factor loadings within the full-scale single-factor solution (<.30), and tested this reduced model. According to all fit parameters, final reduced 18-item single-factor model has shown satisfactory fit, with an average factor loading of .360.

Table 3
Parameters obtained in CFAs[12]

model		no. items per factor	χ^2 (*df*)	χ^2/df	*TLI*	*CFI*	*RMSEA* (CI)
1.	Full-scale single-factor model	36	752.1 (594)**	1.266	.730	.746	.032 (.024-.039)
2.	Full-scale 2-factor model	19/17	698.0 (593)**	1.177	.821	.831	.026 (.017-.034)
3.	3-factor affect model (Harkenss et al. (2005))	8/16/12	742.6 (591)**	1.256	.740	.756	.031 (.024-.038)
4.	2-factor affect model (Konrath et al. (2014))[4]	6/9	109.7 (89)	1.232	.876	.895	.030 (.000-.047)
5.	Single-factor reduced model (Konrath et al. (2014))	17	135.4 (119)	1.129	.923	.932	.022 (.000-.039)
6.	Single-factor reduced model (Olderbak et al. (2015))	10	44.0 (35)	1.257	.857	.888	.032 (.000-.058)
7.	Single-factor reduced model	18	148.5 (135)	1.100	.953	.958	.020 (.000-.037)

Note. χ^2 – chi-square, *df* – degrees of freedom, *RMSEA* – Root Mean Square Error of Approximation, *TLI* – Tucker-Lewis index, *CFI* – Comparative fit index; *RMSEA* ≤ 0.06, *CFI* ≥ 0.95, *TLI* ≥ 0.95 (Hu & Bentler, 1999)

Psychometric properties were again calculated for the single-factor 18-item version of the *RMET*. Results have shown that item sampling adequacy achieved a more satisfying level (*KMO* = .797) ranging from .680 to .887 for

12 Since Konrath et al. (2014) reported only the target word (not item number) for both two-factor and reduced version, and since three of the target words used appear twice in the test, we iteratively tested all combinations of aforementioned items in order to get to the best set of items as indicated by fit parameters. The results of two Konrath et al. (2014) models presented in table 3 and Appendix B are based on the best fitting models including given items.

individual items. Reliability of individual items ranged between .093 and .247, with overall internal consistency remaining at the fair level despite the exclusion of half of the initial item pool (α = .728). Likewise, homogeneity of the 18-item version of the instrument was improved as well (*H1* = .129; *H5* = .677) achieving more satisfying level. Consequentially, the range of internal validity indices for the individual items in 18-item short version was improved – corrected item-total correlations were ranging from .336 to .608, while corrected correlations with principal component extracted from the scale ranged from .356 to .564. In terms of items' content, i.e. stimuli gender and emotional valence of target words (based on the classification of Harkenss et al. (2005)), the final version resulted in ten female stimuli and eight male stimuli, with five negative (1 male, 4 female stimuli), eleven neutral (6 male, 5 female stimuli), and two positive target words (1 male, 1 female stimulus).

In order to demonstrate that the short form of the Serbian version of *RMET* kept its diagnostic power in differentiating between participants with and without ToM deficits, analysis of covariance (ANCOVA) was performed once again, with age and number of years of education taken as covariates. Levene's test has shown equality of error variances across groups [$F_{(3,256)} = 1.650$, $p = .178$]. Results indicated significant main effect of group [$F_{(1,254)} = 24.885$, $p < .001$, $\eta^2_p = .089$], with clinical group performing significantly worse than the group without deficits. Once again, main effect of gender was not observed [$F_{(1,254)} = 0.593$, $p = .442$], while group *x* gender interaction got closer to the threshold of statistical significance [$F_{(1,254)} = 3.474$, $p = .064$, $\eta^2_p = .013$], mainly deriving from the gender differences between participants in the control group $F_{(1,195)} = 8.814$, $p = .003$, $\eta^2_p = .043$].

Discussion

RMET is one of the most popular measures of individual differences in ToM (Baron-Cohen et al., 2001; Olderbak et al., 2015). Despite *RMET's* wide usage and evidence of its validity (Baron-Cohen et al., 1985, 1997, 2001, 2015; Bora et al., 2016; Bora et al., 2009; Ferguson & Austin, 2010; Lai et al., 2012; Lombardo et al., 2007; Losh et al., 2009; Maurage et al., 2011; Olderbak et al., 2015; Schmidt & Zachariae, 2009; Torralva et al., 2009; Vellante et al., 2013; Voracek & Dressler, 2006) there is still limited knowledge of the test's internal psychometric characteristics and its latent structure. Namely, previous studies have shown that *RMET* exhibits poor internal consistency and homogeneity (e.g. Harkness et al., 2010; Khorashad et al., 2015; Olderbak et al., 2015; Prevost et al., 2014; Ragsdale & Foley, 2011; Vellante et al., 2013; Voracek & Dressler, 2006). Furthermore, some authors questioned its unidimensionality and arguing for short-form versions of the test which would resolve the issues of test homogeneity thus enabling more reliable and economic assessment of ToM capabilities (Olderbak et al., 2015), while others favored affect-specific measures derived from the full-scale test (Harkness et al., 2005; Konrath et al., 2014; Maurage et al., 2011). Following this line of research, we aimed to

evaluate internal psychometric characteristics and latent structure of the Serbian adaptation of *RMET* and to propose the most psychometrically sound version of its short-form. Additionally, we aimed to examine this newly adapted *RMET* through examination of its predictive power in differentiating between people with ToM deficits and controls, as well as to replicate typically observed gender differences in the test's performance (e.g. Baron-Cohen et al., 2015; Baron-Cohen et al., 2005; Baron-Cohen et al., 2001; Khorashad et al., 2015; Schiffer et al., 2013; Vellante et al., 2013).

Item analysis has shown that the majority of items of the Serbian adaptation of *RMET* behave in a similar manner regarding their difficulty comparing to other *RMET* adaptations, as well as the original version of the instrument. Namely, the amount of individual item's deviation from difficulty measures provided in previous studies can be considered negligible, especially bearing in mind a wide range of individual item's difficulties documented in previous studies. Contrasting Serbian version of *RMET* to other adaptations and original version of the instrument revealed that the Serbian version significantly deviates only from the Persian one.

Item analysis of *RMET* has shown that test has a number of items with the unbalanced frequency of selection of foils within a number of items. Similar results were obtained in previous studies using this instrument (e.g. Baron-Cohen et al., 2001; Fernández-Abascal, et al., 2013; Girli, 2014; Khorashad et al., 2015; Prevost et al., 2014; Vellante et al., 2013). Namely, a number of items have shown to contain foils that are relatively poor distractors, whose improvement should, in our opinion, be considered for the second revision of the test. Additionally, distribution of scores has shown to be severely skewed and elongated despite the fair representation of the population which is considered to have ToM deficits thus questioning assumptions of normal distribution of this measure.

Results of the full-test psychometric analysis have shown that Serbian version of *RMET* overall has fair psychometric properties. Bearing in mind that previous studies reported on a wide variability in *RMET* internal consistencies, typically falling in the range from .40 to 70 (Harkness et al., 2010; Khorashad et al., 2015; Prevost et al., 2014; Ragsdale & Foley, 2011; Vellante et al., 2013; Voracek & Dressler, 2006), the Serbian version of *RMET* can be considered fairly reliable, compared to other adaptations (e.g. Girli, 2014; Khorashad et al., 2015; Prevost et al., 2014; Vellante et al., 2013). On the other hand, item sampling adequacy indicated lower representativeness of items for the measurement of ToM construct. Similarly, homogeneity parameters indicated to a relatively small amount of commonality between items indicating more than a single source of variance underlying the test's performance. Consequently, results of EFA have shown that two extracted factors accounted only about 12% of the *RMET's* variance. Additionally, these factors seem to be difficult to interpret in a meaningful way, i.e. by means of abilities recruited in the detection of affect-specific mental states presented in the items.

The fact that the test designed for measurement of the unitary construct of ToM exhibited low homogeneity, the issue that has been raised by the previous

studies as well (e.g. Olderbak et al., 2015), served us for examination of the latent structure of Serbian version of *RMET* throughout testing several models based on previous literature. Similarly, as previous studies have shown (e.g. Olderbak et al., 2015; Vellante et al., 2013) full-scale single-factor model exhibited poor fit according to most of the CFA parameters used. Affect-specific three-factor solution (Harkness et al., 2005) could not account for performance on the test resulting in structural validity which was suggested by previous studies as well (Olderbak et al., 2015). The same was true for the affect-specific two-factor model (Konrath et al., 2014), and two-factor model obtained in EFA within this study.

On the other hand, reduced single-factor models based on short-forms of the *RMET* proposed in previous studies (Konrath et al., 2014; Olderbak et al., 2015) exhibited much better structural validity indicating that the optimal solution for increasing *RMET's* structural validity is to eliminate items deviating from unitary ability measured, therefore pointing to the fact that current *RMET's* setting and item pool doesn't have a potential to detect any affect-specific ability on a latent level (if there is such) that would account for the performance on affect-specific content in a meaningful way.

Following the results of the item analysis and the full-scale single-factor loadings, 18-item short-form of the instrument assessing ToM has been proposed. Eighteen-item *RMET* has shown satisfactory internal psychometric properties and latent structure which is in line with theoretical expectations of a single trait underlying ToM abilities captured by this instrument. By means of the items selected, the 18-item version of *RMET* closely corresponds to those suggested by Olderbak et al. (2015) and Konrath et al. (2014), since it includes 70% of the first, and 65% of the items from the latter scale thus indicating concordance between Serbian version and other short-forms of the test.

Finally, both complete and short versions of the Serbian adaptation of *RMET* have shown satisfactory diagnostic validity in differentiating between the participants that are supposed to have ToM-specific deficits and unimpaired controls (Bora et al., 2016; Bora et al., 2009). On the other hand, typically observed "female superiority" (see Baron-Cohen et al., 2015; Baron-Cohen et al., 2005; Baron-Cohen et al., 2001; Khorashad et al., 2015; Schiffer et al., 2013; Vellante et al., 2013) in the performance on *RMET* was not obtained on a whole sample. Several previous studies pointed to the absence of gender differences on *RMET* as well (see Girli, 2014; Olderbak et al., 2015; Baron-Cohen et al., 2015). However, trend-level interaction between participants' group and gender, which derives from gender differences in the control group is directly comparable with those obtained and elaborated by Baron-Cohen and collaborators (see Baron-Cohen et al., 2015; Baron-Cohen et al., 2005).

Conclusion

Overall Serbian adaptation of *RMET* has demonstrated fair psychometric properties and satisfactory correspondence to both original version and other adaptations of the instrument. The proposed short version of the test has

shown satisfactory latent structure that supports the premise of the unitary object of measurement, i.e. general ToM abilities. Additionally, the instrument demonstrated a satisfactory level of validity in differentiating between persons with ToM deficits and unimpaired controls. However, future studies should further address and provide additional evidence on the construct and predictive validity of this test using alternative measures of ToM capabilities on diverse groups of entities sampled both from general and clinical populations.

References

Adams, Jr. R. B., Rule, N. O., Franklin, Jr. R. G., Wang, E., Stevenson, M. T., Yoshikawa,S., ... Ambady, N. (2009). Cross-cultural reading the mind in the eyes: an fMRI investigation. *Journal of Cognitive Neuroscience, 22*, 97–108. doi:10.1162/jocn.2009.21187

Addington, J., Penn, D., Woods, S. W., Addington, D, & Perkins, D. O. (2008). Facial affect recognition in individuals at clinical high risk for psychosis. *The British Journal of Psychiatry, 192*, 67–68. doi:10.1192/bjp.bp.107.039784.

Astington, J. W. (2003). Sometimes necessary, never sufficient: False belief understanding and social competence. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 13–38). New York: Psychology Press.

Baron-Cohen, S., Bowen, D. C., Holt, R. J., Allison, C., Auyeung, B., Lombardo, M. V., ... Lai, M. (2015). The "Reading the Mind in the Eyes" test: complete absence of typical sex difference in ~400 men and women with autism. *PLoS ONE, 10(8)*, e0136521. doi:10.1371/journal.pone.0136521

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*, 813–822.

Baron-Cohen, S., Knickmeyer, R. C., & Belmonte, M. K. (2005). Sex differences in the brain: implications for explaining autism. *Science, 310*, 819–823.doi:10.1126/science.1115455

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition, 21*, 37-46. https://doi.org/10.1016/0010-0277(85)90022-8

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*, 241–251. doi:10.1017/S0021963001006643

Bora, E, Bartholomeusz, C., & Pantelis, C. (2016). Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder. *Psychological Medicine, 46*(2), 253-264. doi:10.1017/S0033291715001993.

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: meta-analysis. *Schizophrenia Research. 109*, 1–9. doi:10.1016/j.schres.2008.12.020

Doherty, M. J. (2009). *Theory of Mind. How children understand others' thoughts and feelings*. Hove, East Sussex: Psychology Press.

Emery N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews, 24*, 581-604. doi:10.1016/S0149-7634(00)00025-7

Ferguson, F. J., & Austin, E. J. (2010). Associations of trait and ability emotional intelligence with performance on theory of mind tasks in an adult sample. *Personality and Individual Differences, 49*, 414–418. doi:10.1016/j.paid.2010.04.009

Fernández-Abascal, E. G., Cabello, R., Fernández-Berrocal, P., & Baron-Cohen, S. (2013). Test–retest reliability of the "Reading the Mind in the Eyes" test: a one-year follow-up study. *Molecular Autism, 4:*33. doi:10.1186/2040-2392-4-33

Frith, U. (2001). Mind blindness and the brain in autism. *Neuron, 32*(6), 969–979. doi:10.1016/S0896-6273(01)00552-9

Frith, C. D. (2007). The social brain? *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 671-678. doi:10.1098/rstb.2006.2003

Girli, A. (2014). Psychometric properties of the Turkish child and adult form of "Reading the Mind in the Eyes Test". *Psychology, 5*, 1321–1337. doi:10.4236/psych.2014.511143

Gordon, R. M. (1986). Folk psychology as simulation. *Mind and Language, 1,* 158–171. http://dx.doi.org/10.1111/j.1468-0017.1986.tb00324.x

Hallerbäck, M. U., Lugnegård, T., Hjärthag, F., & Gillberg, C. (2009). The reading the mind in the eyes test: test-retest reliability of a Swedish version. *Cognitive Neuropsychiatry,* 14, 127–143. doi:10.1080/13546800902901518

Harkness, K. L., Jacobson, J. A., Duong, D., & Sabbagh, M. A. (2010). Mental state decoding in past major depression: effect of sad versus happy mood induction. *Cognition & Emotion, 24*, 497–513. doi:10.1080/02699930902750249

Harkness, K., Sabbagh, M., Jacobson, J., Chowdrey, N., & Chen, T. (2005). Enhanced accuracy of mental state decoding in dysphoric college students. *Cognition & Emotion, 19*, 999–1025. doi:10.1080/02699930541000110

Hawken, E. R., Harkness, K. L., Lazowski, L. K., Summers D., Khoja, N., Gregory, J. G., & Milev R. (2016). The manic phase of bipolar disorder significantly impairs theory of mind decoding. *Psychiatry Research, 239*, 275-280. doi:10.1016/j.psychres.2016.03.043

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

Khorashad, B. S., Baron-Cohen, S., Roshan, G. M., Kazemian, M., Khazai, L., Aghili, Z., ... Afkhamizadeh, M. (2015). The "Reading the Mind in the Eyes" test: investigation of psychometric properties and test–retest reliability of the Persian version. *Journal of Autism and Developmental Disorders, 45*, 2651–2666. doi:10.1007/s10803-015-2427-4

Knežević, G., & Momirović, K. (1996). RTT9G i RTT10G: dva programa za analizu metrijskih karakteristika kompozitnih mernih instrumenata [RTT9G and RTT10G: Two programs for the analysis of metric properties of composite measuring instruments]. In: P. Kostić (Ed.), *Merenje u psihologiji, 2* [Measurement in Psychology, 2] (pp. 35–56). Belgrade: Institute for Criminological and Sociological Research.

Konrath, S., Corneille, O., Bushman, B. J., & Luminet, O. (2014). The relationship between narcissistic exploitativeness, dispositional empathy, and emotion recognition abilities. *Journal of Nonverbal Behavior, 38*, 129–143.doi:10.1007/s10919-013– 0164-y

Kunihira, Y., Senju, A., Dairoku, H., Wakabayashi, A., & Hasegawa, T. (2006). 'Autistic' traits in non-autistic Japanese populations: relationships with personality traits and cognitive ability. *Journal of Autism Development Disorders, 36*, 553–566. doi:10.1007/s10803-006-0094-1

Lai, M. C., Lombardo, M. V., Ruigrok, A. N., Chakrabarti, B., Wheelwright, S. J., Auyeung, B., ... Baron-Cohen, S. (2012). Cognition in males and females with autism: similarities and differences. *PLoS ONE, 7*(10), e47198. doi:10.1371/journal.pone.0047198

Lombardo, M. V., Barnes, J. L., Wheelwright, S. J., Baron-Cohen, S. (2007). Self-referential cognition and empathy in autism. *PLoS ONE, 2*(9), e883. doi:10.1371/journal.pone.0000883

Losh, M., Adolphs, R., Poe, M. D., Couture, S., Penn, D., Baranek, G. T., & Piven, J. (2009). Neuropsychological profile of autism and the broad autism phenotype. *Archives of General Psychiatry, 66*(5), 518–526. doi:10.1001/archgenpsychiatry.2009.34.

Mahy, C. E. V., Moses, L. J., & Pfeifer, J. H. (2014). How and where: Theory-of-mind in the brain. *Developmental Cognitive Neuroscience, 9*, 68–81. https://doi.org/10.1016/j.dcn.2014.01.002

Maurage, P.,Grynberg, D., Noël, X., Joassin, F., Hanak, C., Verbanck, P., ... Philippot, P. (2011). The "Reading the Mind in the Eyes" test as a new way to explore complex emotions decoding in alcohol dependence. *Psychiatry Research, 190*, 375–378. doi:10.1016/j.psychres.2011.06.015

Miu, A. C., Pana, S. E., & Avram, J. (2012). Emotional face processing in neurotypicals with autistic traits: implications for the broad autism phenotype. *Psychiatry Research, 198*, 489–494. doi:10.1016/j.psychres.2012.01.024

Mussen, P. H., & Eisenberg, N. (1977). *Roots of caring, sharing, and helping*. San Francisco: Freeman & Company

Ochsner, K. N. (2008). The social-emotional processing stream: five core constructs and their translational potential for schizophrenia and beyond. *Biological Psychiatry, 64*, 48–61. doi:10.1016/j.biopsych.2008.04.024

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the eyes test: toward a brief form for research and applied settings. *Frontiers in Psychology, 6*, 1503. doi:10.3389/fpsyg.2015.01503

Pfaltz, M. C., McAleese, S., Saladin, A,, Meyer, A. H., Stoecklin, M., Opwis, K., ... Martin-Soelch, C. (2013). The Reading the Mind in the Eyes Test: Test-retest reliability and preliminary psychometric properties of the German version. International Journal of Advances in Psychology, 2(1), e1-e9.

Pijnenborg, G. H. M., Spikman, J. M., Jeronimus, B. F., & Aleman, A. (2013). Insight in schizophrenia: associations with empathy. European Archives of Psychiatry and Clinical Neuroscience, 263(4), 299–307. doi:10.1007/s00406-012-0373-0

Premack, D. G., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?. *Behavioral and Brain Sciences, 1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

Prevost, M., Carrier, M. E., Chowne, G., Zelkowitz, P., Joseph, L., & Gold, I. (2014). The Reading the Mind in the Eyes test: validation of a French version and exploration of cultural variations in a multi-ethnic city. *Cognitive Neuropsychiatry, 19,* 189–204. doi:10.1080/13546805.2013.823859

Ragsdale, G., & Foley, R. A. (2011). A maternal influence on Reading the Mind in the Eyes mediated by executive function: Differential parental influences on full and half-siblings. *PLoS ONE, 6*, e23236. doi:10.1371/journal.pone.0023236

Schiffer, B., Pawliczek, C., Muller, B. W., Gizewski, E. R., & Walter, H. (2013). Why don't men understand women? Altered neural networks for reading the language of male and female eyes. *PLoS ONE, 8*(4), e60278. doi:10.1371/journal.pone.0060278

Schmidt, J. Z., & Zachariae, R.(2009). PTSD and impaired eye expression recognition: a preliminary study. *Journal of Loss and Trauma, 14*, 46–56. doi:10.1080/15325020802537096

Shamay-Tsoory, S. G., Harari, H., Aharon-Peretz, J., & Levkovitz, Y. (2010). The role of the orbitofrontal cortex in affective theory of mind deficits in criminal offenders with psychopathic tendencies. *Cortex*, *46*(5), 668-677. doi:10.1016/j.cortex.2009.04.008

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., & Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, *1*(3–4), 149-166.

Slaughter, V., & Repacholi, B. (2003). *Individual differences in theory of mind: Implications for typical and atypical development*. Hove, UK: Psychology Press. doi:10.4324/9780203488508

Solomon, S., R., & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods, 8*(2), 448-462. doi:10.22237/jmasm/1257034080

Torralva, T., Roca, M., Gleichgerrcht, E., Bekinschtein, T., & Manes, F. (2009). A neuropsychological battery to detect specific executive and social cognitive impairments in early frontotemporal dementia. *Brain, 132*, 1299–1309. doi:10.1093/brain/awp041

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*, 326-354. http://dx.doi.org/10.1080/13546805.2012.721728

Voracek, M., & Dressler, S. G. (2006). 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, 41*, 1481–1491. https://doi.org/10.1016/j.paid.2006.06.009

Wellman H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: the truth about false belief. *Child Development, 72*(3), 655-684. doi:10.1111/1467-8624.00304

Yildirim, E. A., Kasar, M., Güdük, M., Ates, E., Küçüparlak, 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*, 1–8. doi:10.5080/u6500

RECEIVED 04.05.2017.
REVISION RECEIVED 23.08.2017.
ACCEPTED 12.09.2017.

Appendix A

Psychometric properties of individual items calculated by RTT10g macro (Knežević & Momirović, 1996)

	item sampling adequacy	reliability	internal validity	
			H	*B*
i1	.5083	.1818	.1403	.0023
i2	.7106	.2314	.3319	.3404
i3	.4689	.1351	.1312	.1708
i4	.8841	.2990	.5265	.4605
i5	.6214	.1917	.2913	.3091
i6	.6884	.1849	.2673	.3019
i7	.7793	.1620	.3289	.3368
i8	.6882	.3320	.4120	.3819
i9	.7959	.1917	.3502	.3512
i10	.7951	.1871	.3514	.3363
i11	.7556	.2926	.3860	.3372
i12	.7763	.2349	.4245	.3863
i13	.8084	.1917	.4026	.3943
i14	.7130	.2333	.3629	.3689
i15	.8171	.1897	.4026	.3806
i16	.8310	.1936	.3846	.3757
i17	.8224	.3138	.4680	.4231
i18	.8379	.2369	.4442	.4477
i19	.8190	.1773	.3451	.3291
i20	.7169	.2489	.3489	.3771
i21	.5976	.2584	.3050	.3065
i22	.8525	.2314	.4746	.4605
i23	.4441	.2778	.1703	.1872
i24	.5074	.2153	.2614	.2686
i25	.2403	.1799	.0801	.0425
i26	.4160	.1746	.1658	.1970
i27	.7571	.2402	.3966	.3573
i28	.7296	.2218	.3836	.3516
i29	.6895	.1806	.3126	.3135
i30	.6895	.2100	.3601	.3650
i31	.7225	.2500	.3433	.3742
i32	.4315	.2392	.2303	.3000
i33	.7050	.0833	.2178	.2562
i34	.7840	.1925	.3764	.3766
i35	.3562	.1420	.1963	.2237
i36	.6785	.1262	.2591	.2736

Appendix B

Factor loadings for seven models tested

	full-scale single-factor model	full-scale two factor model		Harkenss et al. (2005)			Konrath et al. (2014)		Konrath et al. (2014)	Olderbak et al. (2015)	18-item single-factor reduced model
		F1	F2	positive	neutral	negative	positive	negative	single-factor reduced model	single-factor reduced model	
i1	-.128		-.218	-.081			-.237		-.211		
i2	.286	.355				.263		.196	.205		
i3	.114		.144		.123		.155		.150		
i4	.490		.535		.501		.601		.588		.560
i5	.257		.294			.290		.278	.272		
i6	.226	.288		.282							
i7	.288		.308		.296			.265	.242		
i8	.368		.406		.374				.368	.489	.360
i9	.306	.353			.303				.293	.261	.286
i10	.313		.371		.319						.323
i11	.345		.429			.362		.353	.322		.339
i12	.381		.378		.386			.374	.400	.330	.405
i13	.354	.381			.355		.320		.320		.305
i14	.319	.354				.311		.280	.306	.262	.300
i15	.362	.347			.367				.372	.380	.387
i16	.341	.391		.352			.283		.298		.335
i17	.430		.496			.485		.499	.442		.470
i18	.395	.424			.397						.352
i19	.304		.354		.308					.361	.282
i20	.299	.357		.398							
i21	.259	.321		.361			.185		.193		
i22	.427		.430			.451		.380		.389	.388
i23	.153		.157			.173					
i24	.229		.233		.232					.226	
i25	-.071		-.125	-.029							
i26	.145		.192			.168					
i27	.359		.354			.366		.396	.410		.413
i28	.343	.338			.346						.366
i29	.277	.283			.282						
i30	.313	.314		.365							.277
i31	.291	.407		.386							
i32	.199	.240			.204					.112	
i33	.186	.186			.184						
i34	.335	.341				.355					.325
i35	.169	.179				.164					
i36	.221	.240				.215				.213	