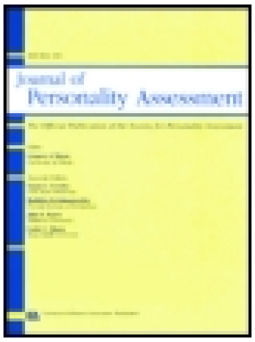

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: http://www.tandfonline.com/loi/hjpa20

An IRT Analysis of the Reading the Mind in the Eyes Test

Jessica E. Black

To cite this article: Jessica E. Black (2018): An IRT Analysis of the Reading the Mind in the Eyes Test, Journal of Personality Assessment, DOI: 10.1080/00223891.2018.1447946

To link to this article: https://doi.org/10.1080/00223891.2018.1447946

Published online: 03 Apr 2018.

JOURNAL OF PERSONALITY ASSESSMENT
https://doi.org/10.1080/00223891.2018.1447946

An IRT Analysis of the Reading the Mind in the Eyes Test

Jessica E. Black

Department of Psychology, University of Oklahoma

ABSTRACT
The Reading the Mind in the Eyes Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), originally designed for use in clinical populations, has been used with increasing frequency as a measure of advanced social cognition in nonclinical samples (e.g., Domes, Heinriches, Michel, Berger, & Herpertz, 2007; Kidd & Castano, 2013; Mar, Oatley, Hirsh, de la Paz, & Peterson, 2006). The purpose of this research was to use item response theory to assess the ability of the RMET to detect differences at the high levels of theory of mind to be expected in neurotypical adults. Results indicate that the RMET is an easy test that fails to discriminate between individuals exhibiting high ability. As such, it is unlikely that it could adequately or reliably capture the expected effects of manipulations designed to boost ability in samples of neurotypical populations. Reported effects and noneffects from such manipulations might reflect noise introduced by inaccurate measurement; a more sensitive instrument is needed to verify the effects of manipulations to enhance theory of mind.

ARTICLE HISTORY
Received 28 August 2017
Revised 8 January 2018

Researchers in different disciplines often use the Reading the Mind in the Eyes Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001) as a performance measure of advanced social cognition (e.g., Barnes & Black, 2015a, 2015b; Domes, Heinrichs, Michel, Berger, & Herpertz, 2007; Hartshorne & Germine, 2015; Kidd & Castano, 2013; Mar, Oatley, Hirsh, de la Paz, & Peterson, 2006; Panero et al., 2016). Although experimental studies using the RMET as an outcome measure have caused an impact in their respective fields, their results have failed to replicate. For example, Domes et al. (2007) reported that administering intranasal oxytocin improved accuracy on difficult RMET items, but Radke and de Bruijn (2015) found no effect of oxytocin. Similarly, Kidd and Castano (2013) found that reading literary fiction (vs. popular fiction, nonfiction, and control) enhanced RMET scores, but two large-scale replication attempts failed to find an effect (Panero et al., 2016; Samur, Tops, & Koole, 2018). Proffered explanations for these nonreplications have ranged from small variations in method and sample characteristics to mechanism or unmeasured third variables (cf. Kidd & Castano, 2017; Panero et al., 2016, 2017; Radke & de Bruijn, 2015). However, here I argue that the primary reason for conflicting results in experimental designs is that RMET is an inadequate measure of social cognition when the assumption is that a treatment will raise ability above baseline in neurotypical adults.

The RMET was designed to capture deficits in social cognition typical of persons with autism spectrum conditions (Baron-Cohen, Wheelwright, Hill, et al., 2001; Vellante et al., 2013). The test consists of 36 items, each of which has an image of the area surrounding the eyes of an adult face (see Figure 1). Participants are asked to choose out of four words the one that best describes the feelings and emotions expressed by the image. A vocabulary list is provided in case participants are unfamiliar with the words, which include terms such as preoccupied, aghast, dispirited, contemplative, and flustered. Internal consistency reliability for the RMET is not consistently reported (it was not reported in the original Baron-Cohen, Wheelwright, Hill, et al. [2001] paper), and when reported, is often low (e.g., Mar et al. [2006] reported $r_\alpha = .60$), as is the mean interitem correlation (Olderbak et al., 2015). Poor internal consistency might reflect the existence of more than one factor; however, the RMET was intended as a unidimensional measure (Baron-Cohen, Wheelwright, Hill, et al., 2001), and studies testing multiple factor structures have concluded that additional factors did not improve fit (e.g., Preti, Vellante, & Petretto, 2017; Vellante et al., 2013; but see Olderbak et al., 2015, for a discussion of issues with fitting factor models to RMET data). Test–retest reliability is somewhat better; although the original paper (Baron-Cohen, Wheelwright, Hill, et al., 2001) does not address it, Vellante et al. (2013)) reported $r_{tt} = .833$.

Originally conceptualized as a measure of "mentalizing" or advanced theory of mind (ToM)—the ability to recognize and interpret mental states, such as intentions and emotions—the RMET was intended for use in clinical populations (Baron-Cohen, Wheelwright, Hill, et al., 2001). It has since been used in nonclinical samples to operationalize various constructs, including ToM (Black & Barnes, 2015, 2015a, 2015b; Kidd & Castano, 2013), mentalizing (Mar et al., 2006; Samur et al., 2018), mind reading (Domes et al., 2007), empathy

CONTACT Jessica E. Black ✉ Jessica.black@ou.edu Department of Psychology, University of Oklahoma, Dale Hall Tower, Room 705, 455 W. Lindsey, Norman, OK 73071.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hjpa.

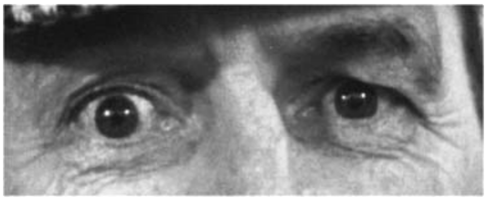

Figure 1. Sample item from the Reading the Mind in the Eyes Test (RMET). The RMET shows participants 36 such images and asks them to choose the best word (of four) to describe what the person in the picture is thinking or feeling. The words for this sample item are jealous, panicked, arrogant, and hateful.

(Djikic, Oatley, & Moldoveanu, 2013), empathic accuracy (Mascaro, Rilling, Negi, & Raison, 2013), and interpersonal sensitivity (Fong, Mullin, & Mar, 2013). In a study designed to contrast ToM with emotion recognition ability, using clinical patients, Oakley, Brewer, Bird, and Catmur (2016) found evidence that the RMET measures emotion recognition rather than ToM. Others suggest that, in neurotypical samples, the RMET measures intelligence, particularly verbal IQ, as well as social cognition (Baker, Peterson, Pulos, & Kirkland, 2014; Peterson & Miller, 2012). Here, I treat the RMET as a measure of ToM (or mentalizing) rather than empathy, and assume that ToM includes understanding others' emotions, recognizing that understanding does not imply any moral or empathic response. Even in high-functioning autism spectrum disorders, there is a dissociation between empathy (which people with high functioning autism do feel) and ToM, or emotion-recognition, as assessed by the RMET (in which the same people show deficits; Montgomery et al., 2016).

The RMET works well to detect the slight deficits in mentalizing typically exhibited by those with Asperger's syndrome or high-functioning autism (Baron-Cohen, Wheelwright, Hill, et al., 2001; Vellante et al., 2013); it might not work as well to discriminate at above-normal levels of ToM. Developing an instrument to effectively test high levels of social cognition is especially challenging given that humans are the most advanced of social primates; superior social-cognitive skills might well have contributed to the evolution of human intelligence (Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007; Hrdy, 2011). Even those who are considered deficient in social-cognitive skills (e.g., autism spectrum disorders) have some understanding of ToM; Montgomery and colleagues reported $M = 20.1$ ($SD = 7.7$) on the RMET for participants with high-functioning autism. This was significantly lower than their participants with Asperger's ($M = 23.5$, $SD = 7.0$), who also have difficulties interpreting emotions (compare this with $M = 26.2$ for neurotypical adults; Baron-Cohen Wheelwright, Hill, et al., 2001). In other words, even populations with social-cognitive deficits are getting more than 50% correct on the 36-item RMET. Yet more problematic for academic researchers, students tend to score even higher on the RMET. Baron-Cohen, Wheelwright, Hill, et al. (2001) reported a mean of 77.7% correct for students, compared with a mean of 72.8% for general population adults. Although some researchers have used adult samples from Amazon.com's Mechanical Turk (MTurk; e.g., Kidd & Castano, 2013; Samur et al., 2018) or adults from the community (Domes et al., 2007; Mascaro et al., 2013), many studies use college students (e.g., Black & Barnes, 2015a, 2015b; Mar et al., 2006; Mar et al., 2009; Radke & de Bruijn, 2015). That said, if the test included difficult items that discriminated well among those with high ToM ability, it would still provide a decent measure for use in nonclinical samples.

The RMET has been used frequently as an outcome measure in experimental studies targeting nonclinical populations; notably, initial studies with large impacts in their respective fields have failed to replicate. For instance, in one of the best known investigations of the effects of oxytocin, Domes et al. (2007) found that participants who had received intranasal oxytocin outperformed those who had received a placebo; the study was double-blind, and used only males. Radke and de Bruijn's (2015) attempt to replicate the original study failed, however, and in a meta-analysis that included studies using both clinical and nonclinical samples, Leppanen, Ng, Tchanturia, and Treasure (2017) found no effect of oxytocin on RMET performance.

A second example comes from the research on the effects of fiction. Kidd and Castano (2013) carried out a series of five studies with results that supported their claim that reading literary fiction enhanced performance on the RMET compared with reading nonfiction, popular fiction, or nothing; Black and Barnes (2015a) replicated Study 1 from Kidd and Castano (2013) in a within-subjects mixed-model design. However, when Samur et al. (2018) carried out a large-scale replication and extension of Kidd and Castano's five studies (including one preregistered exact replication), they found no evidence of an advantage to reading literary fiction over any of the control conditions. Similarly, Panero et al. (2016), combining data from three labs in a mixed model, failed to replicate the effects reported by Kidd and Castano (2013).[1]

The unreliable effect of reading on RMET scores is particularly noteworthy, because of the large samples involved. Debate on the reliability of the effect has been lively (cf. Kidd & Castano, 2017; Panero et al., 2017), in part because there is compelling evidence of a robust correlation between lifelong exposure to fiction and the emotional understanding or ToM captured by RMET scores (e.g., Djikic et al., 2013; Kidd & Castano, 2016; Mumper & Gerrig, 2017; Panero et al., 2016). Of course, it is just as likely that people who are good at mentalizing and enjoy using their social cognitive skills choose to read more fiction as it is that reading fiction improves social cognition. Nevertheless, Kidd and Castano (2017) reanalyzed the data from Panero et al. (2016), and after excluding 245 participants,[2] found an arguably significant difference ($p = .044$, after multiple analyses) between reading literary fiction and popular fiction (with literary fiction enhancing RMET scores), but no difference between reading literary fiction and nonfiction. The apparently unreliable nature of the effects of reading (or oxytocin) on RMET scores might be in part due to differences in

[1] It should be noted that, unlike Kidd and Castano (2013) and Samur et al. (2018), who used MTurk participants only, Panero et al. (2016) used 34 cases (4.3%) from an undergraduate research pool. The rest were recruited on MTurk. Panero et al. used the same stimuli chosen by Kidd and Castano; Samur et al. also used Kidd and Castano's stimuli for the most part, but supplemented it with additional similar texts in one study.

[2] Exclusions were based on reading time and scores of zero on a covariate. Kidd and Castano (2017) also excluded participants from two studies where there was an imbalance across reading conditions not sufficiently explained in Panero et al. (2016).

data exclusion criteria, differences in best practices, and publishing bias, but it might also be due to the instrument itself.

The purpose of this research was to determine the RMET's suitability for use in assessing high levels of social cognition using item response theory (IRT). IRT analysis provides parameter estimates of the level of difficulty of each item, and of item sensitivity to differences in ability at its level, as well as of overall test characteristics. As such, IRT can tell us how difficult the RMET is, how much information each item and the test as a whole provide about the latent trait, and how well it can discriminate between individuals at typical levels of ability. Only two prior studies have applied IRT analyses to the RMET; Carey and Cassels (2013) used IRT to compare the child's version of the RMET (Baron-Cohen, Wheelwright, Spong, Scahill, & Lawson, 2001) to a similar task with open-ended responses, and Preti et al. (2017) performed an IRT analysis of the Italian version of the adult RMET. Carey and Cassels (2013) reported that the child version was most accurate at two standard deviations below the mean, making it appropriate for use in populations with severe deficits only. Preti et al. (2017) confirmed the unidimensional structure of the RMET, and found that the test was most sensitive at below-average ability. This study appears to be the first IRT analysis of the adult English form of the RMET.

Method

Participants

Three labs contributed a total of 591 cases (55.5% female, M age = 30.72, range = 18–66) to this study. Of those who provided data on race or ethnicity, 7.2% were African American, 11.3% were Asian, 6.8% were Hispanic, 77.0% were White, and 2.3% were mixed race or other. All data collections had been approved by the relevant institutional review boards. All participants completed the RMET in an online survey as part of control groups for larger studies ($N = 365$), or as part of a large correlational study conducted online with Amazon.com's MTurk ($N = 226$; these data are part of a paper currently in preparation). Of the 365 cases obtained from control groups in experimental studies, 169 were undergraduates participating through psychology department research pools at separate universities and 196 were recruited on MTurk. Included in these 365 cases are 60 that were used in Black and Barnes (2015b) and 189 that were used in Panero et al. (2016); the remainder come from unpublished studies. Although one of the properties of IRT is that it can provide unbiased estimates of item properties in nonrepresentative samples (Embretson, 1996), it is preferable to use data from participants that have not received any manipulation, and that are representative of the relevant populations (Morizot, Ainsworth, & Reise, 2007). Here, only complete cases were used (only one participant—not included in the sample size just reported—had provided incomplete data and was discarded).

Item response theory

IRT encompasses a range of models that can be used in different psychometric analyses (Hambleton, Swaminathan, & Rogers, 1991; Morizot et al., 2007). All models are based on the *item response curve* (IRC), which illustrates the probability of correct item response as a function of latent trait or ability level. The item difficulty (represented by β) is determined by probability of correct response: In other words, item difficulty is an estimate of the level of the latent trait at which sample participants had a 50% likelihood of answering the item correctly. The simplest IRT model is the Rasch, which assumes that items only vary in difficulty, and that the sensitivity of each item to variations in ability is identical. It is the most restrictive model. A 2PL model assumes that difficulty and sensitivity or discrimination vary across items. Item discrimination (α) is represented by the slope of the IRC at its point of inflection—its difficulty level. Steeper slopes indicate better discrimination, as a person with only slightly less (or more) ability will be much less (or more) likely to get a correct answer. An item with a flat or negative slope provides no information about the latent trait. 3PL models introduce a third parameter that models guessing.

Another important aspect of IRT is the *item information function* (IIF), which illustrates the ability of the item to discriminate between participants at all levels of the latent trait. For example, an easy item will provide the most information about the latent trait at low levels, whereas a difficult one will provide the most information at high levels. The *test information function* (TIF) simply sums the IIFs and shows the precision of the test as a whole: the level of the latent trait where it gives the most information, and the distribution across all levels. Ideally, the TIF will peak at zero, or mean difficulty. In this study, IIFs were used to assess and discard items, and the TIF provides an overall picture of the information provided by the RMET. An important assumption of IRT is that item parameters are invariant across populations and levels of the latent trait. IRT encompasses many different methods and models, but only three primary ones were tested here: the Rasch (1961), or one-parameter model (1PL), the two-parameter model (2PL), and the three-parameter model (3PL; Birnbaum, 1968).

Data analysis

First, the fit of a unidimensional model was verified with confirmatory factor analysis (CFA) using M*plus* (Muthén & Muthén, 1998–2012), which accommodates the use of binary outcome variables (this is in essence a 2PL IRT model). Second, IRTPRO (Cai, Thissen, & du Toit, 2015) and the R (R Core Team, 2015) package ltm (Rizopoulos, 2006) were used to test Rasch, 2PL, and 3PL IRT models using all items (ltm was used to test the Rasch model; IRTPRO was used to test the 2PL and 3PL models; the 2PL model was run in both RStudio and IRTPRO: all parameters matched). Both the ltm package and IRTPRO use maximum likelihood estimation to calculate parameters. Poor items were dropped successively according to item information curves and discrimination parameters. CFA was used to test the factor structure of the shortened tests. Model fit criteria include standardized root mean square residuals (SRMSR), root mean square error of approximation (RMSEA), and the Tucker–Lewis nonnormed fit index (TLI;

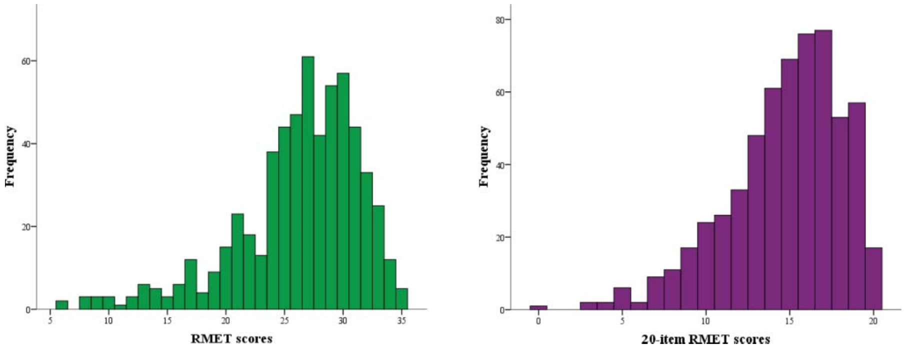

Figure 2. Reading the Mind in the Eyes Test (RMET) score distribution. Both the original and reduced versions of the test are negatively skewed (skew = −11.51 for original; −8.79 for 20-item reduced; $N = 591$).

Kline, 2011). For SRMSR, low values are better; Hu and Bentler (1999) suggested .08 as a cutoff value. RMSEA values with confidence intervals entirely below .05 are considered a good fit (MacCallum, Browne, & Sugawara, 1996). TLI values should be above 0.95 for the fit to be considered good (Hu & Bentler, 1999). Schwarz's (1978) Bayesian information criterion (SBC) was used to compare IRT models (with the same numbers of items): Lower values represent better models.

Results

Descriptive statistics

Mean RMET score was 26.25 ($SD = 5.30$), which was in line with Baron-Cohen, Wheelwright, Hill, et al.'s (2001) reported mean ($M = 26.2$, $SD = 3.60$), $t(711) = 0.11$, $p = .921$, $d = 0.01$. Cronbach's alpha for the complete test was $r_\alpha = .78$; the mean interitem correlation was low, at $r = .095$. The distribution of scores had a strong negative skew (−11.51; see Figure 2); two cases were outliers at > 3.5 *SD* below the mean; nine were outliers at 3.0 *SD* below the mean. Excluding outliers did not correct the negative skew (skew = −8.36); outliers were included in the analyses reported here (Embretson [1996] recommended using a heterogeneous sample). Separate analyses carried out without outliers did not affect the conclusions of the article (results of these analyses are available on request). Women ($M = 26.87$, $SD = 4.50$) scored higher than did men ($M = 25.48$, $SD = 6.08$), $t(467.93) = 3.09$,[3] $p = .002$, but the effect size was not large ($d = 0.26$). See Table 1 for descriptive statistics and mean comparisons.

Table 1. Descriptive statistics and mean comparison for overall scores and by sex and sample for original and reduced Reading the Mind in the Eyes Test.

	N	*M*	*SD*	Min	Max	Median	*t*	*df*	*p*	*d*
36-item										
Overall	591	26.25	5.30	6	35	27				
Female	328	26.87	4.50	8	35	27	3.09	467.9*	.002	0.26
Male	262	25.48	6.08	6	35	27				
MTurk	422	26.54	5.32	6	35	27	2.13	589	.034	0.19
Student	169	25.53	5.19	8	33	27				
20-item										
Overall	591	14.72	3.43	0	20	15				
Female	328	15.09	3.06	3	20	15	2.93	495.1*	.003	0.25
Male	262	14.24	3.81	0	20	15				
MTurk	422	14.73	3.45	0	20	15	0.21	589	.833	0.02
Student	169	14.67	3.40	4	20	15				

[a] Adjust *t*- and *p* values reported due to unequal variances (Levene's test $p < .001$) for mean comparisons (independent samples *t* tests used).

Preliminary model selection

Overall, the hypothesized single-factor model was an acceptable fit for the full 36-item test; although the TLI value was low, RMSEA values were very good, $\chi^2(594) = 891$, $p < .001$, RMSEA = .029, 90% CI [.025, .033], TLI = 0.852; I therefore proceeded to run IRT analyses assuming a unidimensional model. To select the appropriate IRT model (Rasch, 2PL, or 3PL), the models were tested in R. The 2PL model was a significantly better fit than the Rasch model, $\chi^2(35) = 163$, $p < .001$. The 3PL model resulted in a nonpositive definite Hessian matrix, and fit no better than the 2PL, $\chi^2(36) = 4.59$, $p > .99$. The nonpositive definite matrix suggests multicollinearity among parameters and overfitting (Wothke, 1993). As such, and given the advantages of a parsimonious model, the 2PL model was selected, and its item parameters are reported (see Table 1). Because many items had poor discrimination and provided little information, I tested abbreviated versions of the RMET, and then returned to model comparisons with the reduced scale (details later).

Item selection

Initially, item information curves and functions were the basis for discarding items. Table 1 contains descriptive statistics, difficulty (β), and discrimination (α) for each of the original 36 items: Five items had discrimination values of less than 0.50 (graphs of all items are available as supplemental material, Table S.1). Only two items were above average difficulty: Items 10 ($\beta = 0.01$) and 17 ($\beta = 0.15$). As can be seen in Figure 3, the TIF peaks around −1.5; most of the information occurs below the mean (precisely 65.78% lies between 4 *SD* below the mean and zero; only 21.45% lies between mean ability and +4 *SD*). Items were discarded one at a time, based on poor discrimination. The

[3] Adjusted values used to account for unequal variances (Levene's test: $p < .001$).

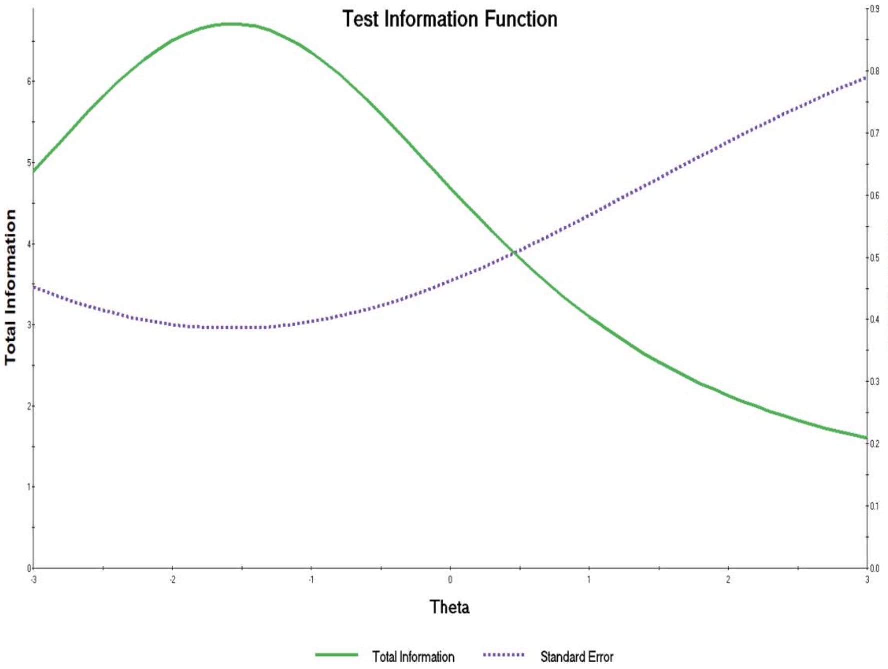

Figure 3. Test information function for original 36-item test. The dotted line shows the standard error. Peak information occurs around 1.5 *SD* below mean ability level (theta).

Table 2. Descriptive statistics and discrimination, difficulty, and item fit statistics for the two-parameter (2PL) model.

Item	Proportion correct	α	*SE*)	β	*SE*	χ^2	*p*
1	.69	0.50	0.11	−1.64	0.37	28.36	.202
2	.72	0.22	0.10	−4.26	2.00	39.09	.019
3	.82	0.47	0.12	−3.43	0.88	17.90	.713
4	.76	0.64	0.12	−1.95	0.35	28.84	.149
5	.89	0.74	0.15	−3.10	0.56	26.36	.193
6	.79	0.56	0.12	−2.56	0.53	26.91	.259
7	.65	0.53	0.11	−1.26	0.29	33.07	.061
8	.86	0.91	0.15	−2.28	0.32	24.72	.259
9	.82	1.00	0.15	−1.82	0.23	20.67	.542
10	.50	0.78	0.12	0.01	0.11	21.41	.375
11	.73	0.87	0.13	−1.34	0.19	23.27	.275
12	.79	1.03	0.15	−1.56	0.19	28.57	.124
13	.75	1.06	0.15	−1.24	0.16	36.48	.013
14	.77	0.98	0.14	−1.42	0.19	44.90	.003
15	.77	1.25	0.16	−1.21	0.14	25.29	.190
16	.70	0.85	0.13	−1.15	0.18	43.03	.005
17	.48	0.46	0.11	0.15	0.18	32.17	.074
18	.68	0.74	0.12	−1.09	0.19	30.36	.085
19	.66	0.65	0.11	−1.08	0.21	22.55	.429
20	.85	0.98	0.15	−2.07	0.27	23.10	.398
21	.89	0.67	0.15	−3.43	0.70	23.09	.233
22	.77	1.21	0.16	−1.27	0.14	27.49	.070
23	.51	0.49	0.11	−0.06	0.17	26.93	.213
24	.80	1.10	0.15	−1.52	0.18	17.50	.682
25	.69	0.54	0.11	−1.58	0.34	28.75	.152
26	.71	0.53	0.11	−1.75	0.37	27.16	.205
27	.77	1.04	0.15	−1.38	0.17	21.42	.375
28	.69	0.74	0.12	−1.20	0.21	17.19	.754
29	.65	0.57	0.11	−1.14	0.25	34.94	.053
30	.88	1.25	0.18	−2.02	0.22	29.87	.072
31	.63	0.30	0.10	−2.26	0.80	27.36	.240
32	.79	0.69	0.13	−2.10	0.36	46.28	.002
33	.72	0.78	0.12	−1.39	0.22	26.32	.285
34	.61	0.96	0.13	−0.52	0.11	26.98	.135
35	.55	0.54	0.11	−0.38	0.17	26.93	.213
36	.90	1.31	0.20	−2.11	0.23	18.96	.526

Note. Proportion correct = proportion of the respondents who gave the correct answer; α = discrimination; $\boldsymbol{\beta}$ = difficulty; *SE* = standard error. Item fit tested with $S - \chi^2$ (a Pearson χ^2 fit statistic; Orlando & Thissen, 2003).

first to be discarded were Items 2 and 31, both of which had alpha values below .31 (see Table 2 and Figure 4).

Items were then discarded one at a time, in order of lowest discrimination, until all items had discrimination values of $\alpha >$ .60,[4] leaving 25 items. This model fit the data better (see Table 3 for results of CFA for all models). Item fit was then referenced, and items with $S - \chi^2$ fit statistic (Orlando & Thissen, 2003) *p* values of $p < .05$ were discarded successively until all items fit (*ps* > .05). The final reduced test had 20 items, all with discrimination values of $\alpha \geq .70$ (see Table 4 for statistics). Both the 25- and 20-item reduced tests were very easy, and few of the items had acceptable discrimination (see Figure 5 for examples of good items). Figure 2 shows the distribution of the 20-item version; Table 1 reported descriptive statistics (see Supplemental Materials for details).

Final model selection

Finally, the three IRT models (Rasch, 2PL, and 3PL) were compared in the reduced 20-item version of the RMET. Again, the 3PL model fit the data no better than the 2PL model: likelihood ratio test, $\chi^2(20) = 6.35$, $p = .998$. The 2PL model fit better than did the Rasch model, $\chi^2(19) = 37.55$, $p = .007$. The reduced version is also preferable to the original in terms of model fit (see Table 3) and item characteristics (Table 4). However, like the original, the reduced version is only informative about the latent variable at low levels of ability (see Figure 6). At greater than average ability, neither of the versions is sensitive enough to

[4] It should be noted that this value ($\alpha > .60$) was arbitrary and does not indicate good discrimination. A major weakness of the RMET is that the items in general do not discriminate well between similar latent trait levels, especially at average to above-average ability.

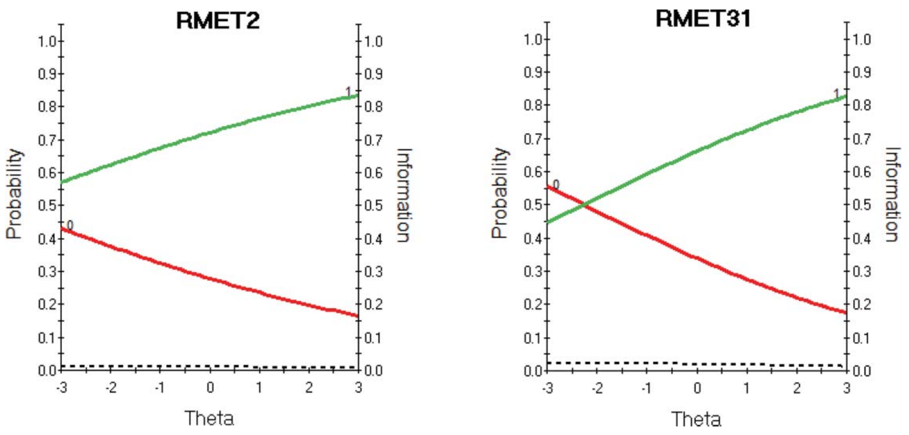

Figure 4. The worst items in terms of discrimination. They were also very easy. The top trace lines (green) show the probability of a correct response; the bottom solid (red) lines show the probability of an incorrect response. The dotted lines at the bottom are the item information function; neither item provides information about latent ability (theta). An individual with 3 *SD* below mean ability had more than a 45% chance of getting Item 31 correct; a person with 3 *SD* above mean ability was not even twice as likely to select the correct response.

discriminate between levels of ToM. In other words, in a neurotypical adult population, people with more than mean ToM ability are likely to have similar scores, making it very difficult, for example, to distinguish between someone whose true ability is +3 *SD* and someone whose ability is only +1 *SD*. All item graphs for the original and 20-item models are available as supplemental material.

Discussion

In line with past research (Carey & Cassels, 2013; Preti et al., 2017), the RMET appears to be an instrument that is best suited for testing low-ability populations. These data suggest that it measures a single construct (see also Preti et al., 2017; Vellante et al., 2013), particularly if the shorter versions suggested here are used, but further testing would be needed to confirm the unidimensional structure. Results support a 2PL model, which estimates item difficulty and discrimination. The most challenging items had difficulty levels around mean latent trait ability. In our sample of neurotypical adults, most items had poor discrimination, making the RMET insensitive to slight differences in ToM at any level, but particularly above the mean. As such, although the RMET might be an excellent instrument for its intended use—detection of deficits in mentalizing typical of individuals with autism spectrum conditions (Baron-Cohen, Wheelwright, Hill, et al., 2001), it is not a good outcome measure if the intention is to differentiate individuals with normal to high ToM ability. Discarding the poorest items, although it might improve test function to some extent, does not alter the fact that the RMET is an easy test with limited ability to distinguish between individuals at similarly high levels of the latent trait. The 20-item version offers a shorter alternative for researchers, but does not comprise an adequate measure for identifying stimuli-induced improvements at high levels of social cognition.

Table 3. Model fit data for full and reduced Reading the Mind in the Eyes Test, all single-factor unidimensional models.

Items	α	Par	χ^2	*df*	*p*	SRMSR	RMSEA	PCLOSE	TLI
36	.78	36	891	594	< .001	0.083	.029 [.025, .033]	1.000	0.852
25	.78	25	342	275	.004	0.069	.020 [.012, .027]	1.000	0.957
20	.73	20	200	170	.058	0.066	.017 [< .001, .026]	1.000	0.971

Note. $N = 591$; α = internal consistency reliability; Par = parameters in model; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation; PCLOSE = one-sided probability of close fit (RMSEA = .05); TLI = Tucker–Lewis Index.

The implications for the research on the immediate effects of intranasal oxytocin or reading fiction are clear: The jury is still out. On the one hand, it is impressive that anyone ever found an effect with the RMET (e.g., Black & Barnes, 2015a, 2015b; Domes et al., 2007; Kidd & Castano, 2013); with a more sensitive measure, we might well find large effect sizes that support inferences of clear, meaningful benefits attached to reading fiction, oxytocin, or both. The item characteristics of the RMET causing noise in the measurement model might have contributed to the failures to replicate (Leppanen et al., 2017; Panero et al., 2016; Radke & de Bruijn, 2015; Samur et al., 2018). On the other hand, experiments using a more appropriate instrument are needed before conclusions can be drawn either way. When it comes to oxytocin, experiments using

Table 4. Discrimination, difficulty, and item fit statistics for reduced, 20-item Reading the Mind in the Eyes Test.

Item	α	*SE*	β	*SE*	χ^2	*p*
4	0.76	0.14	−1.69	0.28	14.50	.415
5	0.78	0.17	−2.98	0.56	8.25	.828
7	0.70	0.12	−0.99	0.19	12.17	.515
8	1.00	0.17	−2.14	0.30	9.02	.830
9	0.95	0.16	−1.88	0.26	9.02	.772
10	0.96	0.14	0.01	0.10	8.21	.769
11	0.89	0.14	−1.33	0.20	5.93	.968
15	1.39	0.19	−1.14	0.13	8.72	.649
16	0.81	0.13	−1.20	0.20	9.01	.831
18	0.70	0.13	−1.16	0.22	17.58	.226
19	0.81	0.13	−0.91	0.16	18.14	.200
24	1.08	0.17	−1.54	0.19	3.79	.987
27	0.96	0.15	−1.47	0.20	9.01	.831
28	1.46	0.23	−1.97	0.21	8.21	.830
30	0.76	0.14	−1.69	0.28	14.50	.415
32	0.78	0.17	−2.98	0.56	8.25	.828
33	0.70	0.12	−0.99	0.19	12.17	.515
34	1.00	0.17	−2.14	0.30	9.02	.830
35	0.95	0.16	−1.88	0.26	9.02	.772
36	0.96	0.14	0.01	0.10	8.21	.769

Note. α = discrimination; $\boldsymbol{\beta}$ = difficulty; *SE* = standard error. Item fit tested with $S-\chi^2$ (a Pearson χ^2 fit statistic; Orlando & Thissen, 2003).

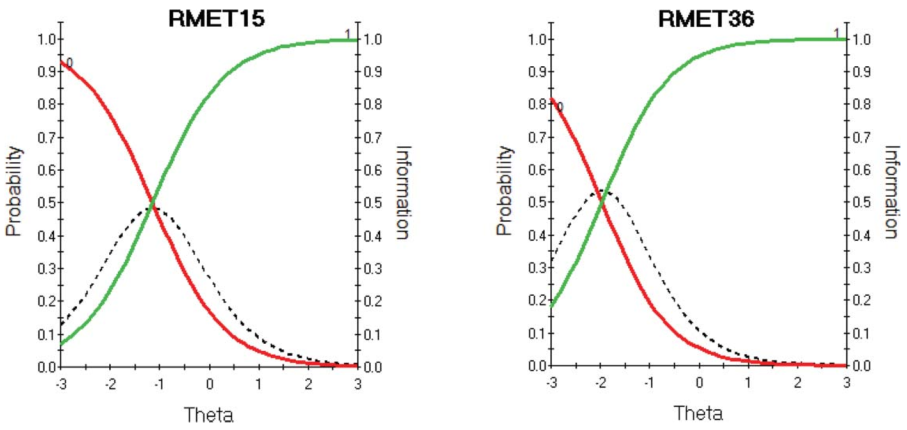

Figure 5. Two of the best items (these graphs are from the 20-item reduced version). Although both items are easy, they are relatively sensitive, and provide information about the latent trait.

different samples (e.g., children, clinical) and outcome measures support the positive association with ToM (e.g., Davis et al., 2013; Wu & Su, 2015). Although there is less experimental evidence for the effect of reading literary fiction on social cognition using alternate measures (but see Pino & Mazza, 2016), correlational studies have shown a more robust positive relationship between RMET scores and reading exposure (Kidd & Castano, 2016; Mumper & Gerrig, 2017; Panero et al., 2016). This study found the negative skew that should be expected in RMET scores (assuming samples of primarily neurotypical adults). Indeed, Panero et al. (2016) and Vellante et al. (2013) reported correcting a negative skew, and Black and Barnes (2015b) reported transforming the variable to meet assumptions of normality (but not how). However, other authors do not address the issue (e.g., Djikic et al., 2013; Kidd & Castano, 2013, 2016; Mar et al., 2006). Although transforming a negatively skewed variable can meet assumptions of normality for statistical purposes, the restriction of range caused by ceiling effects in a measure of performance is likely to cause attenuation in correlations (Lord & Novick, 1968). As such, it could well be that the reported associations between lifetime exposure to reading and RMET scores underestimate the true relationship between reading and ToM ability.

In short, an instrument that accurately measures advanced social cognition in nonclinical samples is needed for experiments that aim to test whether a given manipulation (e.g., intranasal oxytocin or types of narrative) enhances ToM or related constructs. It could be that the RMET can be adapted to be more challenging and more sensitive to changes in latent trait ability (see Carey & Cassels, 2013). Further IRT analyses of the current version are also recommended; in this study, the sample size for undergraduates made differential item testing between these and MTurk participants unfeasible, and it would be informative to test item characteristics in a clinical sample. Differential item testing across groups such as gender and sample source (e.g., MTurk vs. undergraduates) is recommended. What is more, these analyses and conclusions rest on the assumption that the trait—ToM or mentalizing ability—is normally distributed in the population. Given the unique social-cognitive ability of humans (e.g., Churchland, 2011; Herrmann et al., 2007), with most normal individuals possessing high levels of the latent trait, it could be that the population distribution is negatively skewed; future investigations using techniques such as factor mixed modeling (see Hallquist & Wright, 2014) would inform the current discussion.

That said, this study provides important initial information about the suitability of the RMET and the inferences we can

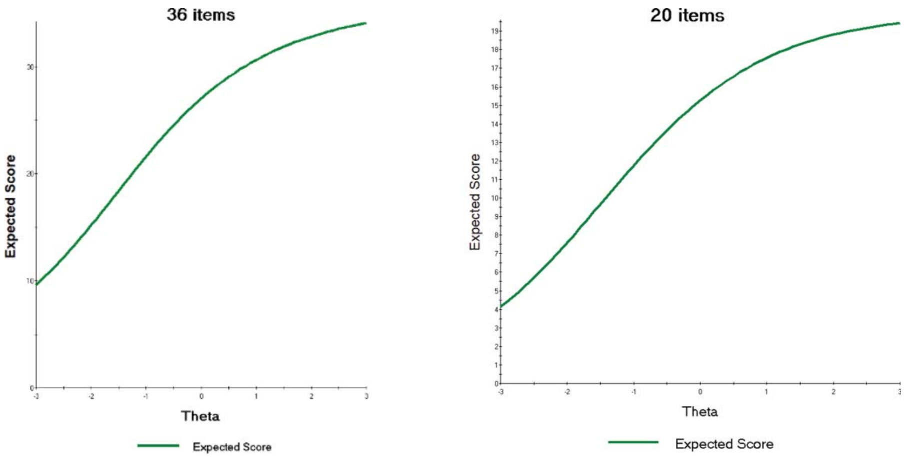

Figure 6. Test characteristic curves for the original (left) and reduced version of the Reading the Mind in the Eyes Test. In both cases there is very little discrimination above mean latent ability (theta = 0).

draw from studies that have used it as an outcome measure of social cognition in nonclinical samples. It is important to note that the failure of the RMET to discriminate at high levels of ability in this study does not reflect on its effectiveness in identifying the deficits in social cognition to be expected in the clinical samples for which it was designed (persons with autism spectrum conditions). Rather, the results of the IRT analyses reported here should be taken as a precaution to researchers who wish to use an instrument designed for clinical samples to assess average to high levels of ability. Future research is needed to develop a valid measure for assessing differences in advanced ToM in neurotypical adults.

Acknowledgments

I would like to thank Deena Skolnick-Weisberg and M. E. Panero for contributing data for this project, and Robert Terry for statistical advice.

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. doi:10.1016/j.intell.2014.03.001

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 Psychiatry*, *42*, 241–251. doi:10.1111/1469-7610.00715

Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., & Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger syndrome. *Developmental and Learning Disorders*, *5*, 47–78.

Black, J. E., & Barnes, J. L. (2015a). The effects of reading material on social and non-social cognition. *Poetics*, *52*, 32–43. doi:10.1016/j.poetic.2015.07.001

Black, J. E., & Barnes, J. L. (2015b). Fiction and social cognition: The effect of viewing award-winning television dramas on theory of mind. *Psychology of Aesthetics, Creativity, and the Arts*, *9*, 423–429. doi:10.1037/aca0000031

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Cai, L., Thissen, D., & du Toit, S. H. C. (2015). *IRTPRO for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International.

Carey, J. M., & Cassels, T. G. (2013). Comparing two forms of a childhood perspective-taking measure using CFA and IRT. *Psychological Assessment*, *25*, 879–892. doi:10.1037/a0032641

Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton, NJ: Princeton University Press.

Davis, M. C., Lee, J., Horan, W. P., Clarke, A. D., McGee, M. R., Green, M. F., & Marder, S. R. (2013). Effects of single dose intranasal oxytocin on social cognition in schizophrenia. *Schizophrenia Research*, *147*, 393–397. doi:10.1016/j.schres.2013.04.023

Djikic, M., Oatley, K., & Moldoveanu, M. C. (2013). Reading other minds: Effects of literature on empathy. *Scientific Study of Literature*, *3*, 28–47. doi:10.1075/ssol.3.1.06dji

Domes, G., Heinrichs, M., Michel, A., Berger, C., & Herpertz, S. C. (2007). Oxytocin improves "mind-reading" in humans. *Biological Psychiatry*, *61*, 731–733. doi:10.1016/j.biopsych.2006.07.015

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*, 341–349. doi:10.1037/1040-3590.8.4.341

Fong, K., Mullin, J. B., & Mar, R. A. (2013). What you read matters: The role of fiction genre in predicting interpersonal sensitivity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*, 370. doi:10.1037/a0034084

Hallquist, M. N., & Wright, A. G. C. (2014). Mixture modeling methods for the assessment of normal and abnormal personality: Part I. Cross-sectional models. *Journal of Personality Assessment*, *96*, 256–268. doi:10.1080/00223891.2013.845201

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science*, *26*, 433–443. doi:10.1177/0956797614567339

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*, 1360–1366. doi:10.1126/science.1146282

Hrdy, S. B. (2011). *Mothers and others: The evolutionary origins of mutual understanding*. Cambridge, MA: Belknap.

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

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*, 377–380. doi:10.1126/science.1239918

Kidd, D. C., & Castano, E. (2016). Different stories: How levels of familiarity with literary and genre fiction relate to mentalizing. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000069

Kidd, D. C., & Castano, E. (2017). Panero et al. (2016): Failure to replicate methods caused the failure to replicate results. *Journal of Personality and Social Psychology*, *112*, e1–e4. doi:10.1037/pspa0000072

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Leppanen, J., Ng, K. W., Tchanturia, K., & Treasure, J. (2017). Meta-analysis of the effects of intranasal oxytocin on interpretation and expression of emotions. *Neuroscience and Biobehavioral Reviews*, *78*, 125–144. doi:10.1016/j.neubiorev.2017.04.010

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149. doi:10.1037/1082-989X.1.2.130

Mar, R. A., Oatley, K., Hirsh, J., de la Paz, J., & Peterson, J. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality*, *40*, 694–712. doi:10.1016/j.jrp.2005.08.002

Mar, R. A., Oatley, K., & Peterson, J. B. (2009). Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes. *Communications*, *34*, 407–428. doi:10.1515/COMM.2009.025

Mascaro, J. S., Rilling, J. K., Negi, L. T., & Raison, C. L. (2013). Compassion meditation enhances empathic accuracy and related neural activity. *Social Cognitive and Affective Neuroscience*, *8*, 48–55. doi:10.1093/scan/nss095

Montgomery, C. B., Allison, C., Lai, M.-C., Cassidy, S., Langdon, P. E., & Baron-Cohen, S. (2016). Do adults with high functioning autism or Asperger syndrome differ in empathy and emotion recognition? *Journal of Autism and Developmental Disorders*, *46*, 1931–1940. doi:10.1007/s10803-016-2698-4

Morizot, J., Ainsworth, A. T., & Reise, S. P. (2007). Toward modern psychometrics: Application of item response theory models in personality research. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 407–423). New York, NY: Guilford.

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *11*, 109–120. doi:10.1037/aca0000089

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Oakley, B. F. M., Brewer, R., Bird, G., & Catmur, C. (2016). Theory of mind is not theory of emotion: A cautionary note on the Reading the Mind in the Eyes Test. *Journal of Abnormal Psychology*, *125*, 818–823. doi:10.1037/abn0000182

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

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S – X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289–298. doi:10.1177/0146621603027004004

Panero, M. E., Weisberg, D. S., Black, J. E., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2016). Does reading a single passage of literary fiction really improve theory of mind? An attempt at replication. *Journal of Personality and Social Psychology*, *111*, e46–e54. doi:10.1037/pspa0000064

Panero, M. E., Weisberg, D. S., Black, J. E., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2017). No support for the claim that literary fiction uniquely and immediately improves theory of mind: A reply to Kidd and Castano's commentary on Panero, Weisberg, Black, Goldstein, Barnes, Brownell, & Winner (2016). *Journal of Personality and Social Psychology*, *112*, e5–e8. doi:10.1037/pspa0000079

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*, 220. doi:10.3389/fpsyg.2012.00220

Pino, M. C., & Mazza, M. (2016). The use of "literary fiction" to promote mentalizing ability. *PLoS ONE*, *11*, e0160254. doi:10.1371/journal.pone.0160254

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*. Advance online publication. doi:10.1080/13546805.2017.1300091

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Radke, S., & de Bruijn, E. R. A. (2015). Does oxytocin affect mind-reading? A replication study. *Psychoneuroendocrinology, 60*, 75-81. doi:10.1016/j.psyneuen.2015.06.006

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the 4th Berkeley Symposium on Mathematical and Statistical Probability*, *4*, 321–334.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25. doi:10.18637/jss.v017.i05

Samur, D., Tops, M., & Koole, S. L. (2018). Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013). *Cognition and Emotion*, *32*, 130–144.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

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. doi:10.1080/13546805.2012.721728

Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.

Wu, N., & Su, Y. (2015). Oxytocin receptor gene relates to theory of mind and prosocial behavior in children. *Journal of Cognition and Development*, *16*, 302–313. doi:10.1080/15248372.2013.858042