UNIVERSITY of CALIFORNIA PRESS | Collabra

van Kuijk, I., et al. (2018). The Effect of Reading a Short Passage of Literary Fiction on Theory of Mind: A Replication of Kidd and Castano (2013). *Collabra: Psychology,* 4(1): 7. DOI: https://doi.org/10.1525/collabra.117

ORIGINAL RESEARCH REPORT

The Effect of Reading a Short Passage of Literary Fiction on Theory of Mind: A Replication of Kidd and Castano (2013)

Iris van Kuijk, Peter Verkoeijen, Katinka Dijkstra and Rolf A. Zwaan

The results reported by Kidd and Castano (2013) indicated that reading a short passage of literary fiction improves theory of mind (ToM) relative to reading popular fiction. However, when we entered Kidd and Castano's results in a *p*-curve analysis, it turned out that the evidential value of their findings is low. It is good practice to back up a *p*-curve analysis of a single paper with an adequately powered direct replication of at least one of the studies in the *p*-curve analysis. Therefore, we conducted a direct replication of the literary fiction condition and the popular fiction condition from Kidd and Castano's Experiment 5 to scrutinize the effect of reading literary fiction on ToM. The results of this replication were largely consistent with Kidd and Castano's original findings. Furthermore, we conducted a small-scale meta-analysis on the findings of the present study, those of Kidd and Castano and those reported in other published direct replications. The meta-analytic effect of reading literary fiction on ToM was small and non-significant but there was considerable heterogeneity between the included studies. The results of the present study and of the small-scale meta-analysis are discussed in the light of reading-times exclusion criteria as well as reliability and validity of ToM measures.

Keywords: theory of mind; literary fiction; replication; meta-analysis

One of the most remarkable aspects of human beings is that they are able to attribute mental states, such as beliefs, knowledge and emotions to themselves and that they realize that other people also have mental states, which may differ from their own. These abilities are commonly referred to as Theory of Mind (ToM) and typically a distinction is made between inferring and representing others' intentions and beliefs (i.e., cognitive ToM) and detecting and understanding others' emotions (i.e., affective ToM) (e.g., Flavell, 1999; Wellman & Gelman, 1992).

Research has demonstrated that reading narrative fiction is positively related to ToM (e.g., Mar, Oatley, Hirsh, dela Paz, & Peterson, 2006; Mar, Oatley, & Peterson, 2009). Recently, Kidd and Castano (2013) put forward the intriguing hypothesis that one type of literature in particular enhances ToM. Specifically, based on different theories of text processing and text representation (e.g., Bruner, 1986; Miall & Kuiken, 1994), they argue that reading literary fiction should increase affective ToM as compared to reading popular fiction. Because works of literary fiction present readers with interesting and complex characters whose behaviour is often inconsistent with social script, they are encouraged to try to understand these characters' intentions and actions, triggering cognitive processes comparable to those involved in affective ToM. By contrast, works of popular fiction are primarily plot driven instead of character driven and as a result, popular fiction is less likely to evoke affective ToM than literary fiction.

To test their hypothesis, Kidd and Castano (2013) conducted four experiments (i.e., Experiments 2 to 5 in their article) using samples of participants from Amazon's Mechanical Turk online labor market. The procedure was the same in each of the experiments. Participants read one of three short literary fiction texts or they read one of three short popular fiction texts (it should be noted that an additional no-reading condition was included in Experiment 2 and Experiment 5). Subsequently, affective ToM was assessed by measuring each participant's ability to infer the appropriate emotion from images of actors' entire faces (i.e., DANVA2-AF in Experiment 2), to select the emotion expressed in images of only the eyes of an actor (i.e., Reading the Mind's Eye Test (RMET) in Experiments 3 to 5), or to select on the basis of visual and linguistic cues, which of four images a central animation character (called Yoni or John) is thinking of or wants (i.e., the Yoni task, Experiments 3 to 5). In all four experiments, participants

Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, NL

Corresponding author: Peter Verkoeijen (p.p.j.l.verkoeijen@essb.eur.nl)

in the literary fiction condition performed better on the employed affective ToM measure than participants in the popular fiction condition.

The results of Kidd and Castano (2013) are remarkable but there are several reasons to be skeptical about them. For one, the participants in Kidd and Castano's experiments were all adults, presumably from a non-clinical population. It is reasonable to assume that these participants, through social interactions in their daily lives, were al highly experienced in affective ToM. Consequently, it is not clear why reading just a brief excerpt of literary fiction would influence ToM as compared to not reading anything, let alone as compared to reading popular fiction, which may also invoke some ToM mentalizing.

Furthermore, the results of a *p*-curve analysis gave rise to concerns about the evidential value of Kidd and Castano's (2013) original findings. Simonsohn, Nelson and Simons (2014) developed the *p*-curve analysis to examine whether a set of significant findings is indicative of a true effect or merely of selective reporting or even *p*-hacking. A *p*-curve refers to the distribution of significant *p*-values, that is, *p*-values smaller than the commonly used threshold for statistical significance of .05. When the null hypothesis is true (when there is no effect), the *p*-values are uniformly distributed, so *p*-values < .025 are as likely to be observed as .025 < *p*-values < .05. However, when there is a true effect, the *p*-curve is skewed to the right, with small *p*-values being more likely to occur than larger *p*-values. Furthermore, Simonsohn and colleagues (2014) propose that the *p*-curve will be skewed to the left when researchers engage in *p*-hacking, that is, when they engage in questionable research practices (see, Simonsohn, Nelson & Simons, 2011) to obtain significant – and hence publishable – results. The original *p*-curve analysis featured three chi-square tests to assess the evidential value of a set of significant findings, the lack of evidential value or the lack of evidential value and *p*-hacking. The most recent and improved version of the *p*-curve analyses focuses on evidential value (see Simonsohn, Simmons, & Nelson, 2015, see also http://www.p-curve.com/ for additional information about the *p*-curve analysis).

To perform a *p*-curve analysis on a set of significant findings a number of requirements should be met: (1) the *p*-values should pertain to the hypothesis of interest, (2) the *p*-values should be independent and (3) the *p*-values should follow a uniform distribution under the null hypothesis. When these requirements are met, the relevant test statistics should be entered in the *p*-curve app (see http://www.p-curve.com/app4/ for the version on December 1st 2017). Subsequently, these test statistics are transformed into *p*-values, which are used to compute a combination of tests consisting of a binomial test and two z-tests, one for the complete *p*-curve and one for the half *p*-curve. The outcomes of this combination test are first used to test the null hypothesis of no effect. If the half *p*-curve test is significant at the .05 level or both the half and the full *p*-curve test show a $p < .1$, the null hypothesis will be rejected. The conclusion is then that the *p*-curve is skewed to the right and hence that the results contain evidential value. If the null hypothesis is not rejected, the second step is to test the hypothesis that the *p*-curve is flatter than you would expect under a true effect examined with low-powered (i.e., a power of .33) studies. If the combination test reveals a $p < .05$ for the full *p*-curve or both the half *p*-curve and binomial test are $p < .1$, then the hypothesis is rejected. This means that the *p*-curve is actually flatter than would one would expect under low power indicating that the evidential value of the findings is inadequate or absent.

When we entered the *F*-values from the critical literary-fiction vs. popular fiction comparison (i.e., Experiment 2: $F(1, 69) = 3.71$; Experiment 3: $F(1, 65) = 4.07$; Experiment 4: $F(1, 68) = 4.39$; Experiment 5: $F(1, 221) = 6.20$) in the *p*-curve applet, we found no indication of evidential value (null of no effect, test for right-skewedness: binomial test, $p = .875$, full *p*-curve, $Z = 1.11$, $p = .8666$, half *p*-curve, $Z = 0.1$, $p = .5405$). In addition, the outcomes of the tests of the null for 33% power, revealed the following results, binomial test, $p = .2043$, full *p*-curve, $Z = -1.92$, $p = .0274$, half *p*-curve, $Z = 0.7$, $p = .7581$. Consistent with the decision rule presented above, we conclude that evidential value of the set of is inadequate or absent because the full *p*-curve gives a *p*-value smaller than .05.

The present study

Although our *p*-curve analysis of Kidd and Castano's (2013) findings casts doubts on the strength of the empirical evidence reported in their paper, Simonsohn, Nelson and Simmons (2014) recommend to always back-up a *p*-curve analysis of a single paper with an adequately powered direct replication of at least one of the studies in the *p*-curve analysis (see http://richarddmorey.org/content/Psynom17/pcurve/#/ for another reason to not rely exclusively on the results of a *p*-curve analysis). Therefore, we conducted a direct replication of the literary fiction condition and the popular fiction condition from Kidd and Castano's Experiment 5 to scrutinize the effect of reading literary fiction on ToM. We focused on Experiment 5 of the original study for several reasons. First, this experiment was conducted with Amazon Mechanical Turk participants and sampling from this pool was the only way for us to conduct a large-scale replication efficiently. In addition, the analyses of the original Experiment 5 were the most complete in the set of five studies as they included a range of important covariates. Because the inclusion of relevant covariates (i.e., variables correlated with the outcome measure) increases the statistical power for the central literary-fiction vs. popular-fiction comparison, we used the same analytic approach in our replication. We attempted to replicate the original Experiment 5 in 2014, and subsequently other direct replications were published (Panero et al., 2016; Samur, Tops, & Koole, 2017). In the discussion, we will reflect on the results of these papers and we will relate them to our own findings and those of Kidd and Castano (2013).

Method

Disclosure statement

We also used the data from the present experiment to support another study published in a Dutch journal [i.e., Dijkstra, Verkoeijen, Van Kuijk, Yee Chow, Bakker &

Zwaan (2015). Leidt het lezen van literaire fictie tot meer empathie? Een replicatiestudie. *De Psycholoog, 50 (10)*, 10–21]. In that study, we applied different exclusion criteria and we used a slightly different analytic approach than in the current study. These exclusion criteria are listed on the Open Science Framework page of the present experiment (https://osf.io/b64mj, see the *Exlusion criteria clarification* Word document). In the SPSS file on the Open Science Framework page named "De Psycholoog data", the filter variable indicates the cases we selected based on the exclusion criteria from the previously mentioned paper. When we analyzed these selected cases in the same manner as we did in the current study (run the "De Psycholoog" syntax on the OSF page for the outcomes of these analyses), we found no effect of type of fiction (literary vs. popular) on ToM measures; an important issue that we will address in the discussion of the present paper.

Ethics statement

In Dutch legislation, the law on medical-scientific research on humans (Wet Medisch Wetenschappelijk Onderzoek met mensen; WMO) serves to protect people from medical maltreatment and experimentation. The WMO applies to research in which people are submitted to a medical/physical intervention or to research in which a certain mode of behavior is imposed on people. According to the WMO, approval from an ethics committee is not required for certain behavioral studies, such as the present experiment (note that it is almost always required for studies involving a medical/physical intervention). For these studies, psychological scientists employed at the Erasmus University Rotterdam are allowed to decide themselves whether they want to consult the Ethics Committee Psychology (ECP). This committee fulfills an advisory role: its members evaluate whether formal approval of a Medical Ethical Committee is required. We concluded that a formal advice of the ECP was not necessary for this experiment because: (1) there is no deceit in the procedure; (2) participants take part on a voluntary basis; (3) the experimental procedure is noninvasive; (4) participants are not expected to experience harm by taking part in the study; (5) participants receive a payment proportionate to the task and time investment at hand; and (6) the results are analyzed and reported in an anonymized fashion.

Thus, we did not ask the ECP for a formal written approval waiver, but we will obtain one if needed.

Participants and design

For the present experiment, we sampled from the same population as Kidd & Castano (2013) in their experiments. Specifically, a total of 558 participants were recruited using Amazon's Mechanical Turk and they were rewarded $1.50 for their participation. Participants were informed about the procedure, the required time investment, and the anonymity of the data analysis. However, we did not obtain written informed consent.

The experimental procedures for the three stories in the literary fiction condition and the three stories the popular fiction condition were posted as separate Human Intelligence Tasks (HITS) on the Mechanical Turk platform. The description of the experimental procedure was the same for each text/HIT. Participants took part in one of the experimental conditions by selecting a HIT. The ToM outcome measures of interest were the scores on the RMET test, and the scores on the first-order and second-order affective and cognitive Yoni task (see the materials section below for detailed information about these five measures).

We used the same exclusion criteria as Kidd & Castano (2013). We excluded the data from participants who did not complete the experiment ($n = 73$), second data entries from participants who had participated in the experiment before (i.e., the second entry could refer to the exact same text version or a different text version, $n = 23$), and data from non-native English speakers ($n = 15$). We also discarded participants with an average reading time of less than 30 seconds per page (page length was standardized; $n = 36$). Outliers (3.5 *SD* below the mean) were removed for the Reading the Mind in the Eyes Test ($n = 1$), critical Yoni trials ($n = 2$) and control Yoni trials ($n = 6$). In addition, participants with an extremely high score (3.5 *SD* above the mean) on the Author Recognition Test foils ($n = 7$) and on average reading time per page were removed ($n = 2$). Our final sample consisted of 393 participants (261 female), who ranged in age from 18 to 81 ($M = 37.3$, $Sd = 13.1$). Of them, 217 participated in the popular fiction condition and 176 in the literary fiction condition.

Materials

We used the same materials as Kidd and Castano (2013) did in their Experiment 5. Below, we provide detailed information about these materials.

Texts. We used the same six texts (three popular fiction, three literary fiction) as in the fifth experiment of Kidd & Castano (2013). The literary fiction texts were: *The VanderCook* by Alice Mattinson, *Corrie* by Alice Munro, and *Uncle Rock* by Dagoberto Gilb. The popular fiction texts included *Jane* by Mary Roberts Rineheart, *Too Many Have Lived* by Dashiell Hammett, and *Space Jockey* by Robert Heinlein. The texts were not truncated and consisted of 2708 to 8318 words. They were divided into 2 to 7 pages and each page was approximately 1200 words long. The literary fiction texts were shorter on average ($M = 4713$ words, $Sd = 739$) than the popular fiction texts ($M = 6814$, $Sd = 1305$).

Reading the Mind in the Eyes Test (RMET). The RMET measures accuracy in emotion perception and is an indication of affective ToM. The test shows 36 still pictures presented one by one to a participant. Each picture presents a person's eye region, accompanied by four possible emotional states (Baron-Cohen, Wheelwright, Hill, Raste & Plumb, 2001). One of the emotional states is the correct answer (see **Figure 1**). For each picture, participants are instructed to select the correct emotional state. If they do not know, they should guess which of the alternatives is correct. The minimum score on the RMET is 0 and the maximum score is 36. People with Asperger syndrome or high functioning autism score significantly

Figure 1: An example of a RMET item. Answer options for this item were: *irritated, comforting, bored,* and *playful.* For this item, *playful* is the right answer.

lower than normal individuals on the RMET (Baron-Cohen et al., 2001). Cronbach's alpha in our sample was sufficient for comparing group means (α = .68, n = 393). However, other studies on the RMET indicate that the test is not homogenous and has a poor internal consistency (Olderbak et al., 2015).

Yoni Task. The Yoni task assesses cognitive and affective ToM by measuring the ability to infer mental states from a combination of verbal, eye gaze and facial expression cues (Shamay-Tsoory & Aharon-Peretz, 2007). The test consists of 64 trials that are presented one-by-one. In each trial, a cartoon outlined face called "John" is presented in the middle of the screen, either looking straight or gazing towards one of four corners of the screen (see **Figure 2**).

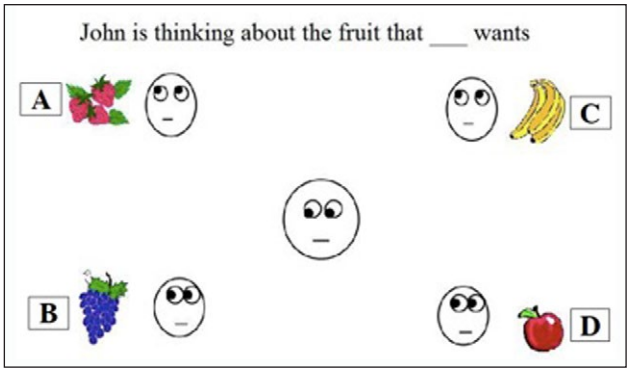

Figure 2: An example of a second-order cognitive Yoni trial. In this case, B is the right answer.

Each corner depicts an object, sometimes accompanied by another cartoon face. Participants are instructed to finish the given sentence (e.g., "John loves ___"); the right answer corresponds to the figure in one of the four corners. If they do not know the answer, they should guess.

There are two experimental conditions: the affective and cognitive condition. These conditions assess affective and cognitive ToM and consist of 24 trials each. Scores on each of the two conditions can range between 0 and 24. The control condition consists of 16 filler trials to check if participants understand the task and avoid responding automatically to eye gaze. The minimum score is 0 and the maximum score is 16. Experimental trials can be further subdivided in first- and second-order trials, based on difficulty level. Each condition (affective and cognitive) consists of 12 first- and 12 second-order items. In the first-order trials, participants have to make inferences of John's thoughts, feelings or physical attributes (e.g.: "John is thinking of ___"). The more difficult second-order trials require understanding of the thoughts, feelings or physical attributes of John regarding the emotions, thoughts or physical attributes of the other four faces (e.g.: "John loves the fruit that ___ loves"). In our sample, the Cronbach's alpha for the affective first-order Yoni task (α = .37, n = 389) and the affective second-order Yoni task (α = .44, n = 383) is insufficient. Furthermore, the Cronbach's alpha for the cognitive first-order Yoni task is poor (α = .50, n = 385), but for the second-order cognitive Yoni task sufficient (α = .70, n = 379). There are other studies that report on psychometric properties of the test, but these used small samples and did not provide any evidence that the test can reliably distinguish healthy subjects from each other (i.e., Shamay-Tsoory, 2008; Shamay-Tsoory & Aharon-Peretz, 2007; Shamay-Tsoory, Harari, Aharon-Peretz, & Levkovitz, 2010).

Author Recognition Test (ART). The ART is an indirect measure of reading habits. It indicates how familiar participants are with fiction authors (Acheson, Wells & MacDonald, 2008; Stanovich & West, 1989). The ART circumvents the problem of social desirability associated with self-report reading questionnaires and is a valid indicator of reading volume (Moore & Gordon, 2015). It consists of a list with 130 names and participants are asked to check off the ones they recognize as authors. However, half of the names on the list are fake authors, so called 'foils'. Marking a foil is associated with a penalty of losing one point on the test. By providing the participants with this information, guessing is discouraged. The Cronbach's alpha of the ART is good (Mol & Bus, 2011). In our sample, the Cronbach's alpha is even excellent (α = .97, n = 393).

Positive and Negative Affect Scale (PANAS). The PANAS is a measure of the two primary dimensions of mood: positive and negative affect (Watson, Clark & Tellegen, 1988). Each of the two mood dimensions consists of a list of 10 words that describe feelings or emotions (e.g., *interested*, *irritable, excited*). Each emotion is rated on a 5-point Likert scale (1 = *not at all*, 5 = *very much*). The positive and negative affect scores each range between 10 and 50. The reliability scores for both scales are moderate to high (Leue & Lange, 2011; Watson et al., 1988). In our sample, Cronbach's alpha scores are excellent for the PANAS negative scale (α = .92, n = 393) and good for the PANAS positive scale (α = .88, n = 393).

Current happiness and sadness. We included two single-item questions assessing current happiness and sadness, which were rated on a 7-point Likert scale (1 = *not at all*, 7 = *very much*).

Transportation Scale. The transportation scale measures to which extent the reader is absorbed into the text (Green & Brock, 2000). It has been suggested that transportation is the key mechanism by which narratives can affect beliefs and empathy (Bal & Veltkamp, 2013). The scale consists of 11 statements that are all presented on the same page, each measured on a 7-point scale (1 = *not at all*, 7 = *very much*). An example statement is: "While I was reading the narrative, I could easily picture the events in it taking place". Three of the items are reverse-scored. The minimum score on the transportation scale is 11

and the maximum score is 77. The Cronbach's alpha of the transportation scale is decent in other studies (Green & Brock, 2000) as well as in our own sample (α = .82, n = 393).

Questions about the text. To measure reader satisfaction, we asked participants whether they thought the text they read was enjoyable and whether they believed it was an example of good literature. These two items were rated on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*).

Perceived Awareness of the Research Hypothesis (PARH). We added five questions at the end of the experiment to measure participant's awareness of the research hypothesis (Rubin, Paolini & Crisp, 2010). The PARH consists of 4 items rated on a 7-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). A sample question is: "I knew what the researchers were investigating with this research". Furthermore, we included a fifth, open question that asked people to write down what they thought was the purpose of the study.

Demographic questions. We included a set of demographic questions, which asked participants to provide information about the age, gender, ethnicity and education level of the participants.

Procedure

The experiment was presented online using the online survey tool Qualtrics. There were six versions of the experiment, corresponding to the six different texts used in this experiment. All other characteristics of the six experiments remained the same (e.g., title and description of the posting, compensation, and order and content of the remaining tasks). Participants were instructed to participate only if they had not completed one of the other versions of the experiment before (they were removed from analysis if they ignored this instruction). The first task was to read the texts. Participants were unaware that reading times per page were recorded. After reading the text, they completed in the following order the RMET, the Yoni task, the PANAS, two items assessing current happiness and sadness, the ART, the Transportation Scale, questions about their satisfaction with the text, demographic questions, and the PARH. Participants were not able to go back in the experiment to change their responses. The exact procedure of our experiment is available on the Open Science Framework at https://osf.io/b64mj/.

Results

Descriptive statistics background variables. Chi-square tests showed that condition (i.e., literary fiction vs popular fiction) was not associated with gender, $\chi^2(1) = 1.725$, $p = .189$, and neither were highest level of education, $\chi^2(5) = 2.283$, $p = .809$, nor self-reported environmental disturbances during the administration of the experiment, $\chi^2(5) = 5.402$, $p = .369$. The mean age in years was similar in the literary fiction condition (M = 36.52, Sd = 12.88) and the popular fiction condition (M = 37.94, Sd = 13.31), $t(391) = 1.070$, $p = .285$. Furthermore, participants in the literary fiction condition enjoyed reading the text (M = 4.57, Sd = 1.94) more than participants in the popular fiction condition (M = 3.90, Sd = 1.90), $t(391) = 3.416$, $p = .001$. In addition, participants in the literary condition considered their texts to be better examples of literature (M = 3.99, Sd = 1.99) than did participants in the popular fiction condition (M = 3.39, Sd = 1.80), $t(391) = 3.137$, $p = .002$. The mean ART score was similar in the literary fiction condition (M = 21.35, Sd = 14.55) and the popular fiction condition (M = 21.30, Sd = 14.76), $t(391) = .029$, $p = .977$. Consistent with the theoretical assumptions made by Kidd and Castano (2013), the level of transportation was higher in the literary fiction condition (M = 48.11, Sd = 10.59) than in the popular fiction condition (M = 41.65, Sd = 11.47), $t(391) = 5.739$, $p < .001$. The average reading time was longer in the literary fiction condition (M = 263s, Sd = 109) than in the popular fiction condition (M = 226s, Sd = 113), $t(319) = 3.707$, $p < .001$. If the literary fiction condition would show a ToM advantage relative to the popular fiction condition, this reading time difference might be a confound. Therefore, we examined the correlations between reading time and ToM variables, i.e., RMET, Yoni-Cognitive first-order, Yoni-Cognitive second-order, Yoni-affective first-order, and Yoni-affective second-order. These correlations were small and non-significant (largest r = .065, smallest p = .201) with the exception of correlation with the RMET scores, i.e., $r = .172$, $p = .001$). Hence, should we find a ToM advantage of literary fiction over popular fiction on the RMET scores, this might be attributed to mean reading time differences.

Correlations/reliabilities background variables and TOM outcome variables. The correlations between the different variables measured in the present experiment as well as Cronbach's alpha of the measures (if applicable) are presented in the correlation matrix in **Table 1**.

Descriptive statistics ToM outcome variables. Cognitive and affective aspects of ToM were measured using the RMET and the first-order and second-order affective and cognitive Yoni tasks. The relevant descriptive statistics on these five outcome measures are presented in **Table 2** as a function of story and condition.

Inferential statistics ToM outcome variables: RMET scores. For all analyses reported below, we will use $p < .05$ as a threshold for statistical significance. To analyse the RMET scores, we used the same analytic approach as Kidd and Castano (2013) did in their Experiment 5. Specifically, we used a linear regression model with the RMET scores as dependent variable and Age, Gender, Highest level of education, Positive affect, Negative affect, Happiness, Sadness, average time spent on RMET items, the ART scores (centered), condition (literary vs. popular fiction) and the ARTxCondition interaction as predictors.

When controlling for all other variables in the model, age, gender, highest level of education, happiness, sadness, ART scores and average time spent on RMET items were not related to the RMET scores, largest F = 1.204, largest partial η^2 = .003. In addition, the ARTxCondition interaction effect was non-significant, $F < 1$. Positive affect $F(1, 381) = 7.381$, $p = .007$, partial $\eta^2 = .019$, and negative affect $F(1, 381) = 7.842$, $p = .005$, partial $\eta^2 = .020$ were both negatively associated with RMET.

Table 1: Reliabilities of the Different Measures in this Experiment (Diagonal Axis) and Correlations between the Different Measures.

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. RMET	.68	.23**	.22**	.19**	.23**	.08	.28**	.02	−.16**	−.17**	−.09	−.10*	0.09	−.002	.11*	−.07
2. Yoni cog 1st		.50	.20**	.49**	.12*	−.003	.05	.06	−.003	−.02	−.03	.06	.003	−.06	.05	−.03
3. Yoni cog 2nd			.70	.27**	.39**	.09	.11*	.13**	.02	−.09	.01	−.02	−.01	−.04	.08	−.03
4. Yoni aff 1st				.37	.28**	.04	.04	−.06	−.07	.02	−.07	.05	−.10	−.04	−.009	−.10*
5. Yoni aff 2nd					.44	.11	.18**	.009	−.04	−.08	−.02	.02	.003	−.12*	.11*	−.03
6. Yoni control						.50	.10	.06	−.01	−.13*	.09	.03	.01	−.02	−.01	−.06
7. ART							.97	.20**	−.03	−.07	−.05	−.04	.37**	.06	.37**	−.04
8. Transportation								.82	.51**	−.09	.31**	−.06	.20**	.06	.09	.01
9. PANAS POS									.88	−.13**	.62**	−.15**	.24**	.03	.03	.07
10. PANAS NEG										.92	−.32**	.51**	−.19**	−.09	−.01	−.03
11. Happiness											–	−.37**	.10	.11*	.01	.03
12. Sadness												–	−.20**	−.05	−.06	−.01
13. Age													–	.12*	.15**	.004
14. Gender														–	.05	.004
15. Education															–	−.08
16. RMET time																–

Note: * $p < .05$, ** $p < .01$. Diagonal values represent reliability estimates of the different measures as quantified by Cronbach's alpha.

Table 2: Mean Scores (standard deviations between brackets) and the First-order and Second-Order Cognitive and Affective Yoni Tasks and mean RMET scores as a Function of Story and Condition.

Story		Yoni Cognitive		Yoni Affective		RMET
		First order	*Second order*	*First order*	*Second order*	
	n	*M (Sd)*	*M (Sd)*	*M (Sd)*	*M (Sd)*	*M (Sd)*
Corrie (L)	61	11.92 (0.28)	10.20 (1.91)	11.70 (0.56)	10.30 (1.24)	28.23 (4.11)
Uncle Rock (L)	58	11.91 (0.28)	10.28 (2.13)	11.62 (0.72)	10.45 (1.11)	27.43 (3.70)
The_Vandercook (L)	57	11.89 (0.41)	10.47 (1.84)	11.51 (0.74)	10.09 (1.31)	27.18 (3.67)
Total Literary fiction	**176**	**11.91 (0.33)**	**10.31 (1.95)**	**11.61 (0.68)**	**10.28 (1.23)**	**27.62 (3.84)**
Jane (P)	70	11.83 (0.48)	9.77 (1.88)	11.54 (0.69)	9.97 (1.44)	26.27 (4.62)
Space Jockey (P)	85	11.74 (0.73)	9.87 (1.91)	11.55 (0.84)	10.14 (1.37)	25.75 (4.87)
Too Many Have Lived (P)	62	11.74 (0.68)	9.68 (2.14)	11.56 (0.78)	10.21 (1.39)	26.95 (4.47)
Total Popular fiction	**217**	**11.77 (0.64)**	**9.78 (1.96)**	**11.55 (0.78)**	**10.11 (1.39)**	**26.26 (4.68)**

Note: maximum score on each of the Yoni tasks = 12, maximum score on the RMET test = 36.

Most importantly, the effect of condition was significant, $F(1, 381) = 12.275$, $p = .001$, partial $\eta^2 = .031$. The estimated marginal mean RMET score in the literary fiction condition was higher (27.68, 95% Confidence interval (CI) = 27.07; 28.28) than in the popular fiction condition (26.22, 95% CI = 25.68; 26.76). This finding is consistent with the literary fiction advantage that Kidd and Castano (2013) observed in their Experiment 5. They found a mean advantage of literary fiction over popular fiction of 1.25 points on the RMET scale (corrected for other terms in the model) compared to a 1.46 advantage in the present experiment. Furthermore, the size of the condition effect was about twice as large in the present experiment as in the original one.

Because the average text reading time scores may have confounded the literary-fiction advantage on the RMET scores, we checked whether the corrected mean difference between the literary-fiction condition and the popular-fiction condition would change when the average text reading time was included as an extra covariate in the analysis above. This did not turn out to be the case. The corrected mean difference, the *p*-value and the effect were hardly affected by adding this covariate.

Inferential statistics ToM outcome variables: Yoni scores. For the Yoni tasks, we also used the same analytic approach as Kidd and Castano (2013) in their Experiment 5. We conducted a mixed ANCOVA on the outcomes of the Yoni tasks, with Yoni Type (cognitive vs. affective) and Yoni Difficulty (first-order vs. second-order) as within-subjects factors, Condition (literary fiction vs. popular fiction) as between-subjects factor, and ART-scores and the scores on the Yoni control task as covariates. It is not clear from their paper why Kidd and Castano used different covariates in the Yoni analysis than in the RMET analysis. However, to allow for a direct comparison with the results from the original study, we adopted the same – albeit inconsistent – approach as Kidd and Castano. Readers interested in investigating the Yoni dependent measures with the covariates from the RMET analysis, are referred to the dataset on the Open Science Framework page.

The analysis showed that the mean (corrected) performance on the first-order Yoni tasks was higher (11.71, 95% CI = 11.66; 11.77) than on the second-order Yoni tasks (10.12, 95% CI = 9.99; 10.26), $F(1, 389) = 14.429$, $p < .001$, partial $\eta^2 = .036$. In addition, when controlling for all other variables in the model, the ART-scores $F(1, 389) = 8.744$, $p = .003$, partial $\eta^2 = .022$ were positively related to the overall Yoni scores. The interaction between Yoni order and the ART-scores was significant, $F(1, 389) = 7.934$, $p = .005$, partial $\eta^2 = .020$.

The effect of condition was significant with higher corrected mean overall Yoni scores in the literary fiction condition (11.03, 95% CI = 10.91; 11.15) than in the popular fiction condition (10.80, 95% CI = 10.69; 10.91), $F(1, 389) = 7.566$, $p = .006$, partial $\eta^2 = .019$. However, this main effect was qualified by an interaction with Yoni Type, $F(1, 389) = 4.597$, $p = .033$, partial $\eta^2 = .012$. Follow-up univariate ANOVA's with condition, ART-scores and Yoni control task scores as predictors demonstrated that the advantage of literary fiction over popular fiction appeared only on the cognitive Yoni tasks, $F(1, 389) = 9.728$, $p = .002$, partial $\eta^2 = .024$, but not on the affective Yoni tasks, $F(1, 389) = 1.890$, $p = .170$, partial $\eta^2 = .005$.

No other effects in the omnibus analysis were significant, maximum $F = 3.574$, maximum partial $\eta^2 = .009$.

Exploratory analyses non-adjusted ToM variables. In the original study of Kidd and Castano (2013), ToM outcome analyses were statistically controlled for various factors, hence we used the same analytic approach in our study. In this section, we also report the unadjusted, uncorrected RMET and Yoni analyses and check if we still obtain the same results.

The uncorrected ANOVA with RMET as dependent measure revealed that even without covariates, the effect of condition was significant and the effect size comparable to the previously reported ANCOVA analysis $F(1, 391) = 9.632$, $p = .002$, partial $\eta^2 = .024$. The estimated marginal mean RMET score was 1.36 points higher in the literary fiction condition (27.62, 95% CI = 26.98; 28.26) than in the popular fiction condition (26.26, 95% CI = 25.69; 26.84).

We also performed an uncorrected, mixed ANOVA with Yoni Type (cognitive vs. affective) and Yoni Difficulty (first-order vs. second-order) as within-subjects factors and Condition (literary fiction vs. popular fiction) as between-subjects factor. The outcomes strongly resembled the

corrected, ANCOVA analyses. We found a main effect for Difficulty, $F(1, 391) = 568.369$, $p < .001$, partial $\eta^2 = .59$, with higher mean scores on first-order Yoni trials (11.71, 95% CI = 11.66; 11.77) than on second-order Yoni trials (10.12, 95% CI = 9.98; 10.26). Furthermore, just as with the corrected analyses, we found a main effect for Condition, $F(1, 391) = 7.422$, $p = .007$, partial $\eta^2 = .019$. Performance on the Yoni task was higher in the literary fiction condition (11.03, 95% CI = 10.91; 11.15) than in the popular fiction condition (10.80, 95% CI = 10.69; 10.91). This effect was again explained by the interaction with Yoni Type, $F(1, 391) = 4.616$, $p = .032$, partial $\eta^2 = .012$. Follow-up univariate ANOVA's revealed that the difference between the two conditions was only significant for the cognitive Yoni task, $F(1, 391) = 9.662$, $p = .002$, partial $\eta^2 = .024$. Performance on the cognitive Yoni task was higher in the literary fiction condition (22.22, 95% CI = 21.91; 22.54) than in the popular fiction condition (21.55, 95% CI = 21.27; 21.84).

The key results, i.e., the effect fiction type on the RMET scores and the cognitive YONI scores, are presented graphically in **Figures 3** and **4**.

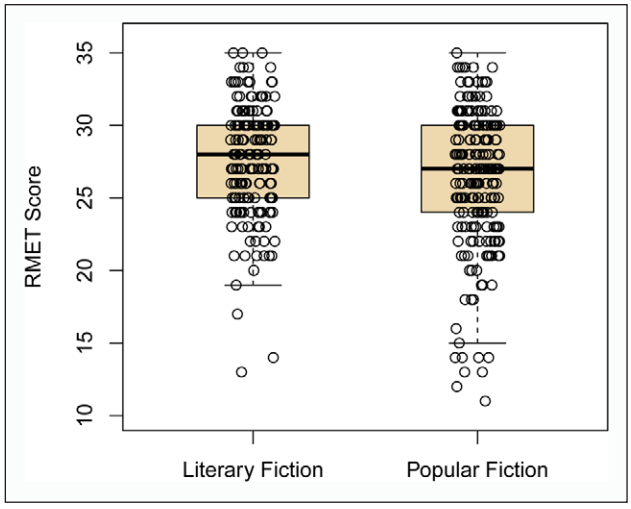

Figure 3: Boxplot and Individual Unadjusted RMET Scores as a function of Condition.

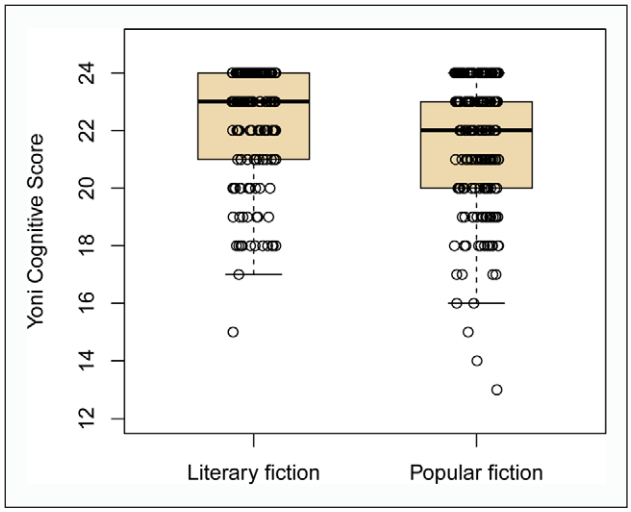

Figure 4: Boxplot and Individual Unadjusted Cognitive Yoni Scores (Collapsed over Difficulty Level) as a function of Condition.

In sum, the uncorrected ToM analyses show similar results compared to the corrected ANCOVA analyses. Thus, it appears that statistically controlling for covariates is not required to obtain significant effects.

Small-scale meta-analysis RMET scores. Apart from the present study, other studies have tried to replicate Kidd and Castano's (2013) literary-fiction advantage using RMET as the dependent variable (i.e., Panero et al., 2016; Samur et al., 2017). We searched for relevant articles using the databases Google Scholar and PsycINFO and the keywords "Kidd and Castano" and "replication". We selected only experimental studies that compared reading literary fiction to reading popular fiction and that used the RMET as outcome variable. Consequently, studies like Black and Barnes (2015), Pino and Mazza (2016) and Experiment 1 of Samur et al. (2017) were excluded. The studies included in the meta-analysis can be found in **Table 3**. To obtain (1) a more precise estimate of the mean RMET difference between the literary-fiction condition and the popular fiction condition and (2) an indication of the variability in mean RMET differences between the six studies, we conducted a small-scale random-effects meta-analysis (see Cumming, 2012). It should be noted that the presented condition means in **Table 3** represent the means adjusted for other terms in the model.

The small-scale meta-analysis showed a non-significant mean RMET difference between the literary-fiction condition and the popular fiction condition (including the original study). i.e., 0.79, 95% CI [–0.05999; 1.6582]. However, there was considerable heterogeneity between studies in terms of the mean RMET differences, $I^2 = 70\%$, $Q(5) = 10.106$, $p = .02$. This is probably due to the fact that the present study and the original study showed comparable literary-fiction advantages, whereas this advantage was absent in the other studies.

Discussion

The goal of the present study was to assess the robustness of Castano and Kidd's (2013) finding that reading only a brief passage of literary fiction enhanced ToM scores as compared to reading popular fiction. Therefore, we conducted a direct replication of Kidd and Castano's Experiment 5. In line with the original finding, we found that when controlling for age, gender, education level, positive affect, negative affect, happiness, sadness, ART scores, and time spent on RMET items, reading literary fiction resulted in higher affective ToM scores (as measured by the RMET) than reading popular fiction. Because participants in the literary fiction condition had longer reading times than those in the popular fiction condition, we checked if the results remained the same when controlling for reading times as well. The results were unaffected. Regarding the Yoni task, Kidd and Castano (2013) found that both cognitive and affective ToM scores were higher in the literary fiction condition. However, our results demonstrated that this advantage of literary fiction over popular fiction only applied to cognitive ToM.

Hence, the results of our direct replication are largely consistent with Kidd and Castano's (2013) findings. However, two other studies that attempted to replicate Kidd and Castano's findings failed, despite the large

samples that were used (Panero et al., 2016; Samur et al., 2017). When we combined the experiments from the original study with the present direct replication and the other published replication in a small-scale meta-analysis, we found a small and non-significant mean RMET advantage of the literary fiction over the popular fiction condition. Furthermore, the studies showed substantial heterogeneity concerning RMET differences between conditions.

The observed heterogeneity may be due to differences in exclusion criteria. For example, Kidd and Castano (2017) criticized the failed replication study of Panero et al. (2016) for not adopting the same exclusion criteria as they did. After correction for these criteria, which involved removing participants with unrealistically short reading times and those that did not take the ART, the mean RMET scores became significantly higher in the literary fiction condition than in the popular fiction condition. Likewise, in their Experiments 3b and 4, Samur and colleagues (2017) appeared to have used a slightly different exclusion criterion than Kidd and Castano did. Specifically, they excluded participants who spent less than 30s per page instead of 30s per page on average that was used in the original study and in the present study. We have some empirical evidence that this difference may matter. In a previous Dutch publication of the current experiment (Dijkstra et al., 2015), we used slightly different exclusion criteria on the exact same data as in the present study, with the most important difference being reading times exclusions (for more information, see https://osf.io/b64mj). Like Samur and colleagues (2017), we excluded participants when they spent less than 30s per page and consistent with Samur and colleagues' results, but inconsistent with the original study and the present study, we failed to find an effect of fiction type on ToM scores. Taken together, it seems that replicating the results of Kidd and Castano (2013) hinges on choosing a particular set of exclusion criteria that *a priori* seem not better than alternatives. In fact, with respect to the studies by Samur and colleagues (2017) and Dijkstra and colleagues (2015), one could argue that a more stringent criterion regarding reading times (i.e., smaller than 30s per page rather than smaller than 30s per page on average) is to be preferred because participants who spent less than 30 seconds on a page did not adhere to the task instruction of reading the entire text carefully.

Apart from the specific exclusion criteria, the diverging findings between the studies in our small-scale meta-analysis may be due to inadequate psychometric properties of the ToM tests. For example, studies indicate that the RMET is not homogenous and typically has a poor internal consistency (Khorashad et al., 2015; Olderbak et al., 2015; Vellante et al., 2012). Although the RMET appeared to have an acceptable internal consistency in our study, Olderbak et al. (2015) warn that measures like Cronbach's alpha may be biased by long tests like the RMET, and that it does not take into account a test's homogeneity. In other words, the reliability of the RMET is questionable. Moreover, the psychometric properties of the Yoni task are not extensively examined. The only available studies we are aware of used small samples and did not provide evidence that the test can discriminate among healthy subjects (i.e., Shamay-Tsoory, 2008; Shamay-Tsoory & Aharon-Peretz, 2007; Shamay-Tsoory et al., 2010). In fact, the discriminative power seems problematic as the Yoni task showed ceiling effects in our study. Furthermore, we also found that the internal consistency of the test was poor. The low reliability of the two ToM tasks may cause relatively high non-systematic variance in these main outcome variables, which in turn may explain the high degree of heterogeneity between studies.

In addition to the issues with the reliability of ToM measures, researchers have expressed concerns about the validity of the RMET test. That is, RMET performance was found to be largely dependent of vocabulary knowledge (Oldenbak et al., 2015; Peterson and Miller, 2012). Assuming that reading literary texts impacts language processing due to style and complexity, RMET differences may reflect language processing differences rather than ToM differences. Compared to popular fiction, reading literary fiction might encourage participants to process the meaning of words, sentences and their relationships more deeply and that might produce ToM differences. Of course, this is mere speculation and more research is needed to flesh out this hypothesis and to test it.

Table 3: Relevant Descriptive Statistics and 95% Confidence Interval of the Mean RMET Difference Between the Literary Fiction Condition and the Popular Fiction Condition for the Studies in the Small-Scale Meta-Analysis.

Study	Literary fiction	Popular fiction	Mean Difference	95% CI Mean Diff.
Kidd & Castano (2013): Experiments 3, 4, 5 combined	*M* = 26.13 *Sd* = 4.10 *n* = 225	*M* = 24.50 *Sd* = 4.92 *n* = 183	1.63	0.7521; 2.5079
Panero et al.,(2016): Experiments 1, 2, 3 combined	*M* = 26.24 *Sd* = 5.74 *n* = 342	*M* = 26.05 *Sd* = 7.01 *n* = 152	0.19	–0.9894; 1.3694
Samur et al., (2017): Experiments 2, 3b, 4 combined	*M* = 27.33 *Sd* = 4.72 *n* = 218	*M* = 27.43 *Sd* = 4.68 *n* = 222	–0.01	–0.9808; 07808
Present study	*M* = 27.62 *Sd* = 3.84 *n* = 176	*M* = 26.26 *Sd* = 4.68 *n* = 217	1.36	0.4976; 2.2224

The current study may have two limitations. First, participants were not randomly assigned to the conditions. We put six versions of the experiment on MTurk simultaneously, corresponding to the six different texts (literary or popular fiction). Although we made sure that the titles, descriptions and compensations of the postings were the same, participants were free to choose the experiment that they wanted to join. It could be possible that the MTurkers discussed the versions with each other, which may have created a bias in participant selection. However, we explicitly asked participants whether they told other participants about the experiment. None of them answered "yes" to this question. Yet even if participants discussed the experiment versions (i.e., if they had not responded truthfully to our question), the total procedure length would have hardly differed between versions because the largest part of the procedure did not consist of reading the text but of taking a series of questionnaires. In addition, demographic variables did not significantly differ between the six versions. Furthermore, the literary fiction texts were on average shorter than the popular fiction texts. Hence, if we assume that motivational factors influence text selection (i.e., that less motivated participants are more likely to select shorter texts) and that motivation is positively related to ToM performance, then our text assignment procedure would actually work against a literary fiction ToM advantage. In other words, in case of a motivation confound, the advantage we observed would be an underestimation of the true (non-confounded) effect.

Another limitation of the current study is that reading time differed between conditions. Participants in the literary fiction condition displayed longer reading times than participants in the popular fiction condition, even though literary fiction stories were on average shorter (M = 4713 words, Sd = 1739) than popular fiction texts (M = 6814 words, Sd = 1305). Furthermore, reading time was positively correlated with RMET performance. Consequently, the ToM advantage of the literary fiction condition could possibly be explained by longer reading times, which indicate more careful reading in this condition. If we were to assume that this reading time difference between conditions reflects non-systematic variance, then we could look at the corrected RMET mean difference between the conditions from our ANCOVA. This difference suggests that even after controlling for reading time differences, the literary fiction condition outperforms the popular fiction condition. However, if the reading time difference is systematic, statistically controlling for covariates does not guarantee a non-confounded comparison (see Miller & Chapman, 2001 for an elaborate and nuanced view on this ANCOVA issue). It may be relevant to note that in the previous publication of the current study (Dijkstra et al., 2015) the reading time difference was not present, nor were there any significant ToM effects. The ToM advantage of literary fiction over popular fiction on the RMET scores may thus be attributed to mean reading time differences. Future studies on this topic may address the role of reading time in enhancing ToM.

The present study largely replicated the findings of Kidd and Castano (2013) and suggests that reading literary fiction, compared to popular fiction, may benefit ToM abilities. However, a small-scale meta-analysis combining the original study with the replications we are aware of indicates that (1) the meta-analytic effect of fiction type on RMET is small and non-significant and (2) there is considerably heterogeneity among the studies. The latter may be due to specific exclusion criteria or problems with the reliability of the RMET scale. Both issues need to be addressed before we can draw strong conclusions about the effect of reading literary fiction on ToM.

Data Accessibility Statement

Examples of the stimuli and presentation materials in the literary fiction condition and popular fiction condition, participant data, and analysis scripts can be found on this paper's project page on https://osf.io/b64mj/.

Acknowledgements

We thank David Kidd and Emanuele Castano for sharing with us valuable information about the materials and the procedure of their original study. Furthermore, we thank Dalya Samur for sharing information about her experiments.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

- Contributed to conception and design: IvK, PV, KD, RZ
- Contributed to acquisition of data: IvK, KD
- Contributed to analysis and interpretation of data: IvK, PV, KD, RZ
- Drafted and/or revised the article: IvK, PV, KD, RZ
- Approved the submitted version for publication: IvK, PV, KD, RZ

Author Information

Rolf A. Zwaan is a Senior Editor at Collabra: Psychology. He was not involved in the peer review of the article.

Peter Verkoeijen is an Editor at Collabra: Psychology. He was not involved in the peer review of the article.

References

Acheson, D. J., Wells, J. B., & **MacDonald, M. C.** (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*, 278–289. DOI: https://doi.org/10.3758/BRM.40.1.278

Bal, P. M., & **Veltkamp, M.** (2013). How does fiction reading influence empathy? An experimental investigation on the role of emotional transportation. *PLoS ONE, 8*, 1–12. DOI: https://doi.org/10.1371/journal.pone.0055341

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.

The Journal of Child Psychology and Psychiatry, 42, 241–251. DOI: https://doi.org/10.1111/1469-7610.00715

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

Bruner, J. (1986). *Actual minds, possible worlds.* Cambridge, Massachusetts: Harvard University Press.

Dijkstra, K., Verkoeijen, P., Van Kuijk, I., Chow, S. Y., Bakker, A., & **Zwaan, R. A.** (2015). Does reading literature result in higher empathy? A replication study. *De Psycholoog, 10*, 10–20.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology, 50*, 21–45. DOI: https://doi.org/10.1146/annurev.psych.50.1.21

Green, M. C., & **Brock, T. C.** (2000). The role of trans portation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*, 701. DOI: https://doi.org/10.1037/0022-3514.79.5.701

Khorashad, B. S., Baron-Cohen, S., Roshan, G. M., Kazemian, M., Khazai, L., Aghili, Z., et al. (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: https://doi.org/10.1007/s10803-015-2427-4

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

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

Leue, A., & **Lange, S.** (2011). Reliability generalization. An examination of the positive affect and negative affect schedule. *Assessment, 18*, 487–501. DOI: https://doi.org/10.1177/1073191110374917

Mar, R. A., Oatley, K., Hirsh, J., Dela Paz, J., & **Peterson, J. B.** (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 Psychology, 40*, 694–712. DOI: https://doi.org/10.1016/j.jrp.2005.08.002

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

Miall, D. S., & **Kuiken, D.** (1994). Foregrounding, defamiliarization, and affect: Response to literary stories. *Poetics, 22*, 389–407. DOI: https://doi.org/10.1016/0304-422X(94)00011-5

Miller, G. A., & **Chapman, J. P.** (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*, 40–48. DOI: https://doi.org/10.1037/0021-843X.110.1.40

Mol, S. E., & **Bus, A. G.** (2011). To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin, 137*, 267–296. DOI: https://doi.org/10.1037/a0021890

Moore, M., & **Gordon, P. C.** (2015). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior Research Methods, 47*, 1095–1109. DOI: https://doi.org/10.3758/s13428-014-0534-3

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: https://doi.org/10.3389/fpsyg.2015.01503

Panero, M. E., Weisberg, D. S., Black, J., 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. DOI: https://doi.org/10.1037/pspa0000064

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: https://doi.org/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: https://doi.org/10.1371/journal.pone.0160254

Rubin, M., Paolini, S., & **Crisp, R. J.** (2010). A processing fluency explanation of bias against migrants. *Journal of Experimental Social Psychology, 46*, 21–28. DOI: https://doi.org/10.1016/j.jesp.2009.09.006

Samur, D., Tops, M., & **Koole, S. L.** (2017). Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013). *Cognition and Emotion, 1*, 1–15. DOI: https://doi.org/10.1080/02699931.2017.1279591

Shamay-Tsoory, S. G. (2008). Recognition of 'fortune of others' emotions in Asperger syndrome and high functioning autism. *Journal of Autism and Developmental Disorders, 38*, 1451–1461. DOI: https://doi.org/10.1007/s10803-007-0515-9

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

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*, 668–677. DOI: https://doi.org/10.1016/j.cortex.2009.04.008

Simonsohn, U., Nelson, L. F., & **Simmons, J. P.** (2014). *P*-curve: A Key to the File Drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547. DOI: https://doi.org/10.1037/a0033242

Stanovich, K. E., & **West, R. F.** (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24*, 402–433. DOI: https://doi.org/10.2307/747605

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., et al. (2012). The "Reading the Mind in the Eyes" test: Systematic review of psychometric properties and a validation study in Italy. *Cognitive Neuropsychiatry, 18*, 1–12. DOI: https://doi.org/10.1080/13546805.2012.721728

Watson, D., Clark, L. A., & **Tellegen, A.** (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*, 1063–1070. DOI: https://doi.org/10.1037/0022-3514.54.6.1063

Wellman, H. M., & **Gelman, S. A.** (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337–375. DOI: https://doi.org/10.1146/annurev.ps.43.020192.002005

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.117.pr

How to cite this article: van Kuijk, I., Verkoeijen, P., Dijkstra, K., & Zwaan, R. A. (2018). The Effect of Reading a Short Passage of Literary Fiction on Theory of Mind: A Replication of Kidd and Castano (2013). *Collabra: Psychology,* 4(1): 7. DOI: https://doi.org/10.1525/collabra.117

Senior Editor: Simine Vazire

Editor: Ed Vul

Submitted: 08 October 2017 **Accepted:** 30 January 2018 **Published:** 27 February 2018

UNIVERSITY of CALIFORNIA PRESS | Collabra

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS