Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013)

Dalya Samur, Mattie Tops and Sander L. Koole

Department of Clinical Psychology, VU University Amsterdam, Amsterdam, The Netherlands

ABSTRACT

Prior experiments indicated that reading literary fiction improves mentalising performance relative to reading popular fiction, non-fiction, or not reading. However, the experiments had relatively small sample sizes and hence low statistical power. To address this limitation, the present authors conducted four high-powered replication experiments (combined \( N = 1006 \)) testing the causal impact of reading literary fiction on mentalising. Relative to the original research, the present experiments used the same literary texts in the reading manipulation; the same mentalising task; and the same kind of participant samples. Moreover, one experiment was pre-registered as a direct replication. In none of the experiments did reading literary fiction have any effect on mentalising relative to control conditions. The results replicate earlier findings that familiarity with fiction is positively correlated with mentalising. Taken together, the present findings call into question whether a single session of reading fiction leads to immediate improvements in mentalising.

One of the most remarkable products of human imagination is literary fiction, the creation of narrative worlds populated by complex characters whose inner lives invite exploration. Literary fiction, perhaps more than other types of texts, creatively engages readers in a discourse and leads them into the subjective world of others (Hakemulder, 2000; Kuiken, Miall, & Sikora, 2004; Mar & Oatley, 2008). Reading literary fiction may thus not only be enriching, but may also be a useful tool in helping people to improve social-emotional skills (Samur et al., 2013). It is therefore of considerable interest to learn more about the psychological effects of reading literary fiction.

Several studies have observed that familiarity with fiction is positively correlated with mentalising performance, that is, people's ability to understand the mental states of others (Dijkic, Oatley, & Moldoveanu, 2013; Mar, Oatley, Hirsh, dela Paz, & Peterson, 2006; Mar, Oatley, & Peterson, 2009). These correlational findings suggest that reading fiction and mentalising may involve overlapping cognitive abilities. In line with this general idea, Kidd and Castano (2013) hypothesised that a single session of reading literary fiction may prime mentalising processes, by leading readers to engage in mind-reading and character construction. To test this hypothesis, Kidd and Castano experimentally manipulated whether or not participants were exposed to literary texts during a single reading session. After this priming manipulation, participants completed various standardised tests of mentalising performance, including the well-established Reading the Mind in the Eyes Test (RMET) (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). In the RMET, people are presented with a series of pictures of the eye region of target persons whose face displays various emotional expressions, such as disappointment, joy, or desire. Participants are asked to indicate for each set of eyes which mental state they believe is most fitting. In five independent experiments, mentalising performance was
significantly higher after their participants had read some literary fiction, compared to when participants had read popular fiction (Experiments 2–5), non-fiction texts (Experiment 1), or no text (Experiments 2 and 5). From these findings, Kidd and Castano concluded that reading literary fiction primes mentalising, thus making people better able to understand others’ mental states.

The findings by Kidd and Castano (2013) could have far-ranging practical implications. For instance, to the extent that these findings are correct, people may be taught to prime themselves with literary fiction in contexts where it is essential for people to understand each other’s mental states, such as negotiation settings (De Dreu, Koole, & Steinel, 2000). Likewise, educators may include literary fiction in teaching curricula and therapists might use literary fiction to help people with mentalising deficits (see Samur et al., 2013). However, before such steps are implemented, it is important to verify the empirical robustness of Kidd and Castano’s findings. Four of the five experiments that were conducted by Kidd and Castano had rather modest sample sizes, varying between 72 and 114 for experiments with 2–3 experimental conditions. The one experiment that did have a larger sample size (N = 356) had the smallest effect size of the entire set of experiments. Thus, stronger statistical evidence is desirable in this paradigm. Indeed, scientific rigour requires researchers to conduct independent replication studies of others’ findings (Koole & Lakens, 2012; Roediger, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). So far, there have been few attempts to extend the results by Kidd and Castano (Black & Barnes, 2015; Bormann & Greitemeyer, 2015). However, none of these studies aimed to stay as close as possible to the original research, such as keeping the same set of texts, procedures, and Amazon Mechanical Turk samples. In view of these considerations, we wanted to try and replicate the original experiments by Kidd and Castano (2013). While we were in the process of publishing our findings, another large replication attempt (N = 792) was published, which found no support for any short-term effect of reading literary fiction on mentalising (Panero et al., 2016). We return to this article in the General Discussion.

In the present article, we report the findings of four replications of Kidd and Castano (2013). Similar to Kidd and Castano, participants were randomly assigned to one of four reading conditions. In one of these conditions, participants read literary fiction. In the other reading conditions, participants read non-fiction (Experiment 1–4), popular fiction (Experiment 2–4), or no text (Experiment 3–4). We then assessed mentalising performance using the RMET. Our replication efforts focused on the RMET, given that this was the main task that was investigated by Kidd and Castano. We conducted our four replication experiments between October 2013 and March 2014. All participants from the present replication experiments were recruited using Amazon’s Mechanical Turk service, as in the original experiments by Kidd and Castano (2013). All four experiments were close replications, in the sense that they followed the Kidd and Castano (2013) procedures as closely as possible. However, the three experiments also included minor procedural modifications for exploratory purposes (e.g. Experiment 3 also included a premeasure of theory of mind). Our fourth experiment omitted any exploratory elements, and was a direct replication of Experiment 3 by Kidd and Castano (2013).

We pre-registered this last experiment in the Open Science Framework (OSF; https://osf.io/bd27z/?view_only=f2e8f4cedde943dca44bfbac73f5939d).

In the literary and popular fiction conditions in our all experiments, we used a combination of short stories and excerpts from novels that are from the original article by Kidd and Castano (2013). We used the excerpts from the novels that were from “the first several pages (8–11)” as it was described in the supplementary materials (Page 4) of the original article (Kidd & Castano, 2013). Only for the non-fiction condition, we introduced three new texts in Experiment 2 and only one text was replaced in Experiment 3.

Based on Kidd and Castano’s research, we expected that RMET scores would be significantly higher after participants had read literary fiction than in any of the control conditions (reading popular fiction, non-fiction, or no reading). Similar to Kidd and Castano, we also measured participants’ familiarity with literature, using an objective test that relies on recognition of the names of published authors (Acheson, Wells, & MacDonald, 2008). Based on prior research (Mar et al., 2006; Mar et al., 2009), we expected that familiarity with literature would be positively associated with performance on the RMET.

Besides replicating Kidd and Castano’s (2013) original research, we also wanted to explore possible individual differences in the effects of reading literary fiction. To this end, we included measures of theoretically relevant individual differences, namely personality questionnaires, in the present studies. These tests are Bermond–Vorst Alexithymia Questionnaire (Vorst
& Bermond, 2001), Transportation Scale (Green & Brock, 2000), the Osnabrück Life Stress Scale (Baumann, Kaschel, & Kuhl, 2005), and Adult Survey of Reading Attitudes (Smith, 1990). The relevant questionnaires were never included in between the given reading condition and the RMET and were included only after reading and the RMET in Experiments 3 and 4. We do not report the effects of these individual differences variables here. Interested readers may contact the first author of this report for more details on the individual difference variables.

**Experiment 1**

In Experiment 1, we compared the effects of reading literary fiction to the effects of reading non-fiction on mentalising performance. This experimental design and all the chosen texts corresponded for both conditions (literary and non-fiction) with the first experiment of Kidd and Castano (2013). Moreover the current experiment had twice as many participants (around 80 per condition) as the original study (around 40 per condition).

**Method**

**Participants and design**

One hundred and sixty participants were recruited using Amazon’s Mechanical Turk service and completed the experiment using the link from Qualtrics Software. They participated for payment ($2.50). In this service, experimenters can approve the performance of participants as sufficient. High approval rate of a participant tends to produce better quality data. Therefore, we followed the common protocol to include participants with more than 95% approval rate. Only in the last two studies, we used additional filters to only include participants from United States. As in the original experiment, we checked the reading times, author recognition guessing rates and the RMET scores for potential outliers (>3.5 SD). The guessing rates for the author recognition test were calculated by the number of the foils selected. We excluded three participants with excessively long reading times and one participant with high guessing rates. There were no outliers detected in the other measure. In the final sample there were 156 participants (69 women, $M_{age} = 33.79$, ages ranged between 18 and 66). One participant did not report his age.

**Materials**

As experimental stimuli, we used six texts (three fiction, three non-fiction) that were all also included in the experiments by Kidd and Castano (2013). The literary fiction texts were “The Runner” by DeLillo (2011), “Blind Date” by Davis (2011) and “Chameleon” by Chekhov (1979). The prose texts were “How the Potato Changed the World” by Mann (2011), “Bamboo Steps Up” by Gandel (2008) and “The Story of the Most Common Bird in the World” by Dunn (2012). Participants were randomly assigned to read one of the six texts.

To measure the mentalising performance, participants completed the RMET, in which they were asked to identify emotions from a set of 36 pictures which only included the eye area (Baron-Cohen et al., 2001). Participants had four options which listed the emotions that they could choose from. We computed scores for RMET as the sum of correctly identified emotions from the presented faces.

Participants proceeded with a task that gauged their familiarity with literature, which is called the author recognition test (Acheson et al., 2008). Specifically, participants were asked to identify the author names in a list with an equal number of compelling foils. There were in total 130 names in the list with 65 identifiable author names and 65 foils. We computed author recognition scores by subtracting the incorrect scores from correct scores and then these results were square root transformed to improve normality in accordance with the original article.

We also included a personality questionnaire, namely the fantasy subscale from Interpersonal Reactivity Index (IRI) (M. H. Davis & Association, 1980), as a check to compare the individual differences in absorption between the conditions. Participants responded to the statements with a 5-item Likert scale, ranging from strongly disagree to strongly agree. It includes 7 items (e.g. I really get involved with the feelings of the characters in a novel).

**Procedure**

After providing informed consent, participants filled out a number of personality questionnaires (as mentioned above). Then, participants were randomly assigned to read one of the six readings listed above. Participants completed RMET and then proceeded to author recognition test. Finally, participants answered other personality questionnaires,
demographic questions and were debriefed about the aim of the research.

**Results and Discussion**

The author recognition test was administered in all the conditions \((M = 22.63, SD = 14.63)\) and the scores did not differ between conditions, \(F(1, 154) = .42, p = .52, \eta^2_p = .00\). The fantasy subscale from IRI was measured in all the conditions and the scores did not differ between the conditions, \(F(1, 154) = .34, p = .56, \eta^2_p = .00\).

We analysed the results using analysis of variance (ANOVA) with condition as between participants factor and author recognition test as covariate. Significance tests for all the present experiments are analysed in IBM SPSS statistics software. The results of these significant tests are displayed in Table 1, all the relevant means and standards deviations are displayed in Table 2.

In the statistical analyses, the higher order variables, such as interaction, have an additive effect on the prediction of main effects (Cohen, Cohen, West, & Aiken, 2003). Therefore, to interpret the main effects, the model should be analysed without the interaction. While, for comparison with a similar table in the paper by Kidd and Castano (2013), Table 1 includes the main effects and interaction from the complete model, the main effects from a separate analysis that excludes the interaction within the model are described in the results section of each experiment.

Replicating prior research (Kidd & Castano, 2013; Mar et al., 2009) there was a strong simple main effect for author recognition, such that higher author recognition was associated with better RMET performance, \(r = .38, F(1, 153) = 25.64, p < .001, \eta^2_p = .14\). The latter finding shows that our assessment of mentalising performance was sufficiently sensitive to detect the association with lifetime exposure to fiction found in previous studies.

We then turned to the effects of the experimental reading manipulation. After reading literary fiction, participants had somewhat higher RMET scores compared to participants who had read non-fiction. However, unexpectedly, this simple main effect was not statistically significant, \(F(1, 153) = .56, p = .45, \eta^2_p = .00\). Thus, our first experiment found no evidence to support Kidd and Castano’s (2013) finding that a single session of reading literary fiction improves mentalising performance.

**Experiment 2**

In Experiment 2, we replicated the design of Experiment 1, while adding a condition in which participants read popular fiction. The latter condition was also used in Experiments 2–5 by Kidd and Castano (2013). This experiment had twice as many participants (around 80 per condition) as the original study (around 40 per condition).

As a minor divergence from the original procedures, we used somewhat different texts in Study 2. We did this to ensure that our findings were not due to some unforeseen idiosyncrasy of the specific texts that participants had to read. Specifically, we picked texts that were used in different experiments by Kidd and Castano for Literary fiction (one from Experiment 2, two from Experiment 5) and Popular fiction (one from Experiment 2 and two from Experiment 5) conditions. For the non-literary fiction condition, we included excerpts that were not previously used by Kidd and Castano (2013); however, we carefully selected the latter texts from the same website that was the source of the non-fiction texts in Study 1. For this, we followed the selection criteria that were used by Kidd and Castano (2013) that the non-fiction texts were mainly focused on a non-human subject, but the chosen texts might include short passages about human subjects.
Method

Participants and design

We recruited 240 participants in the same way as in Experiment 1. We again inspected author recognition guessing rates, the RMET scores and reading times for outliers (>3.5 SD). There were three outliers for the author recognition test, three outliers for low RMET scores and three outliers in reading time. This resulted in a final data set of 231 participants (114 women, $M_{\text{age}} = 36.51$, range 20–70).

Materials

We used nine texts (three fiction, three non-fiction, three popular fiction) in accordance with Kidd and Castano (2013). The fiction texts included “The Tiger’s Wife” by Obreht (2011), “Uncle Rock” by Dagoberto Gilb (Furman, 2012) and “The Vandercook” by Alica Mattinson (Furman, 2012). The non-fiction texts included “How the Chicken Conquered the World” by Adler and Lawler (2012), “Mistletoe: The Evolution of a Christmas Tradition” by Dunn (2011) and “The Venus Flytrap’s Lethal Allure” by Tucker (2010). The popular fiction included excerpts from “Gone Girl” by Gillian Flynn (Flynn, Whelan, & Heyborne, 2012), “Space Jockey” by Robert Heinlein (Hoppenstand, 1998) and “Jane” by Mary Roberts Rinehart (Hoppenstand, 1998). Participants were randomly assigned to read one of the nine texts. Other materials were the same with Experiment 1, namely RMET and the author recognition test.

Procedure

The basic procedure and materials were the same as in Experiment 1, such that participants (1) completed personality questionnaires and mood scale; (2) read the texts, during which we manipulated whether participants read literary texts; (3) completed the RMET; and (4) completed the author recognition test and finally answered some demographic questions. However, two things were different compared to Experiment 1. First, we added a control condition in which participants read a popular fiction text. Second, we replaced all the texts in the fiction and non-fiction conditions with different texts to ensure that the results were not specific to the chosen texts.

Results

The author recognition scores ($M = 22.34$, $SD = 14.76$) did not differ between conditions ($F(2, 228) = .11$, $p = .90$, $\eta^2_p = .00$). The fantasy subscale from IRI was measured for all the conditions, and the scores did not differ between the conditions ($F(2, 228) = 1.55$, $p = .22$, $\eta^2_p = .01$). However, this result showed a trend toward a significant effect. Because we wanted to
make sure that we did not overlook any potentially meaningful effects of the reading manipulation, we explored this nonsignificant trend in the next analyses.

We checked if RMET performance in the literary fiction condition differed from RMET performance in the non-fiction condition, \( F(1, 153) = 3.31, p = .07, \eta^2_p = .02 \). Inspection of Table 2 shows that mentalising was better in the non-fiction condition than in the literary fiction condition. Thus, the effect went in the opposite direction as the findings of Kidd and Castano (2013). Similarly, there was a marginally significant difference in RMET performance in the popular fiction condition as compared to the non-fiction condition, \( F(1, 149) = 4.44, p = .04, \eta^2_p = .03 \), indicating that mentalising was better in the non-fiction condition than in the popular fiction condition. Finally, there was no difference in RMET performance in the literary fiction condition as compared to the popular fiction condition, \( F(1, 151) = .01, p = .92, \eta^2_p = .00 \).

The pairwise comparison results stayed almost the same when the analysis was done without dropping the interaction; literary versus nonfiction condition, \( F(1, 152) = 3.88, p = .05, \eta^2_p = .03 \), popular versus nonfiction condition, \( F(1, 148) = 4.24, p = .04, \eta^2_p = .03 \), and literary versus popular fiction \( F(1, 150) = .01, p = .93, \eta^2_p = .00 \).

**Experiment 3**

In Experiment 3, we used the same design and similar sample size (around 90 participants per condition) as in Experiments 1 and 2, with three notable differences. First, as an additional control, we added a third condition in which we did not present any text to participants. This condition was also used by Kidd and Castano (2013; Experiments 2 and 5). Second, we changed the presentation format of the texts. In Experiments 1 and 2, the stories were typed and presented to the participants on a single page. To rule out that our findings were due to an idiosyncracy in presentation format, we divided the texts according to a standardised page length, a presentation format that corresponds with Kidd and Castano’s original Experiments 3–5 (2013). Third, to clearly demonstrate the effect of the single session, a pre-session baseline RMET and a post-session RMET is required. We therefore measured baseline pre-manipulation RMET performance alongside with the post-manipulation RMET to improve the sensitivity of the experimental design.

We further added two new questions at the end of the experiment. First, we asked for the participant’s native language. Second, we introduced an exploratory question, which was “Did you previously participate any surveys on MTurk that had highly similar content (such as reading fiction, eyes recognition task … et cetera?)”.

**Method**

**Participants and design**

We recruited 366 participants in the same way as in Experiments 1 and 2. Because we now presented the texts in standard page length as Kidd and Castano (2013) did in their Experiments 3–5, we used the criterion for excluding participants based on low reading times (<30s per page) similar to original Experiments 3–5. Based on this criterion, 37 participants were excluded. Applying the same outliers criterion from Experiments 1 and 2 did not result in the removal of any further participants for RMET scores and four participants were excluded because of their high Author Recognition guessing rates. The final data set consisted of 325 participants (171 women; \( M_{age} = 37.31 \), range 19–73).

**Materials**

We selected texts from Experiments 1 and 2 for use in Experiment 3. Nine texts (3 fiction, 3 non-fiction, 3 popular fiction) were used in accordance with Kidd and Castano (2013). The fiction texts included “The Runner” by DeLillo (2011), “Uncle Rock” by Dagoberto Gilb (Furman, 2012) and “The Vandercook” by Alica Mattison (Furman, 2012). The non-fiction texts included “How the Potato Changed the World” by Mann (2011), “How the Chicken Conquered the World” by Adler and Lawler (2012) and “The Story of the Most Common Bird in the World” by Dunn (2011). The popular fiction included excerpts from “Gone Girl” by Gillian Flynn (Flynn et al., 2012), “Space Jockey” by Robert Heinlein (Hoppenstand, 1998) and “Jane” by Mary Roberts Rinehart (Hoppenstand, 1998). Participants were randomly assigned to one of the four experimental conditions (fiction, non-fiction, popular fiction, no text).

We used the same materials as in Experiments 1 and 2. However, we made one important procedural change in the measurement of mentalising performance, the RMET. The RMET was originally developed...
to assess stable individual differences in mentalising abilities (e.g. relating to autism, Baron-Cohen et al., 2001). We therefore wanted to make sure that any experimental effects of reading literary fiction were not overshadowed by chronic individual differences. To this end, we administered a pre-test of 18 items from the RMET before we introduced the reading manipulation. We wanted these 18 items to be representative for the complete test. Therefore based on the relevant literature and the performance scores in Experiments 1 and 2, we balanced the ratio of easy and difficult items similar to the original test (Baron-Cohen et al., 2001). After the reading manipulation, participants completed the full RMET with 36 items. In this manner, we were able to statistically control for any differences in mentalising ability that existed prior to our experimental manipulation.

Procedure

Participants completed the first part of RMET with 18 items and then read the short story followed by the full RMET with 36 items. Then, participants completed the demographic questions, personality questionnaires and finally the mood scale.

Results

The author recognition scores (M = 20.14, SD = 16.31) did not differ between conditions (F(3, 321) = 1.23, p = .30, \( \eta^2 = .01 \)). The fantasy subscale from IRI was measured in all the conditions, and the scores did not differ between the conditions (F(3, 321) = 1.64, p = .18, \( \eta^2 = .02 \)).

We analysed RMET scores using a repeated-measures ANOVA with condition as between participants factor, time (pre-manipulation versus post-manipulation) as within-participant factor and author recognition test as covariate. Because the pre-manipulation score was based on half the number of items that were matched on difficulty with the other half of the items, we multiplied this score by two to make it comparable to the post-manipulation score (see Table 1).

There was a simple main effect of time, F(1, 320) = 16.09, p < .001, \( \eta^2 = .05 \), reflecting that RMET performance decreased from pre-manipulation (M = 26.92; SD = 6.18) to post-manipulation (M = 26.05; SD = 6.34). Author recognition showed an interaction with time, F(1, 320) = 8.36, p < .005, \( \eta^2 = .03 \). The interaction reflects that a higher author recognition score relates to a smaller decrease in RMET performance from pre- to post-manipulation. The crucial interaction between condition and time that would indicate an effect of reading condition on RMET performance was not significant, F(3, 320) = .32, p = .81, \( \eta^2 = .00 \).

Post-manipulation RMET scores were also analysed using ANOVA with condition as between participants factor and author recognition test and pre-manipulation RMET scores as covariate. After controlling for the initial RMET score, there were no significant effects. There was a large main effect for pre-RMET scores (F(1, 316) = 297.31, p < .001, \( \eta^2 = .49 \)) on post-RMET, as expected. The main effect for author recognition task was also significant (F(1, 316) = 47.33, p < .001, \( \eta^2 = .13 \)). The main effect of condition (F(3, 316) = .31, p = .82, \( \eta^2 = .00 \)) and the interaction effect between the condition and the author recognition (F(3, 316) = 1.55, p = .20, \( \eta^2 = .01 \)) were not significant. When we dropped the interaction term from the model, the main effects remained similar for pre-RMET (F(1, 319) = 308.65, p < .001, \( \eta^2 = .49 \)), the author recognition (F(1, 319) = 51.88, p < .001, \( \eta^2 = .14 \)) and condition (F(3, 319) = .29, p = .84, \( \eta^2 = .00 \)).

Additionally, when the pre-RMET is checked using ANOVA with condition as a between participants factor and author recognition test scores as covariate, the main effect of condition remained marginal, F(3, 320) = 2.21, p = .09, \( \eta^2 = .02 \) reflecting that there might be differences between groups already before the reading manipulation (see Table 2).

Perhaps repeating the RMET items influenced the results. To rule out this possibility, we ran the analysis for only the non-repeated RMET items in the post-manipulation, including the pre-manipulation RMET scores as covariate. The results showed a main effect for the pre-manipulation, F(1, 316) = 174.87, p < .001, \( \eta^2 = .36 \), a main effect for Author Recognition, F(1, 316) = 68.77, p < .001, \( \eta^2 = .18 \), no main effect of condition, F(3, 316) = .11, p = .96, \( \eta^2 = .00 \), and no interaction effect, F(3, 319) = .57, p = .64, \( \eta^2 = .01 \).

We also checked for condition by time interactions for the literary fiction condition in comparison to other conditions independently. There were no differences between literary fiction and non-fiction, F(1, 152) = .07, p = .80, \( \eta^2 = .00 \), literary and popular fiction, F(1, 155) = .04, p = .85, \( \eta^2 = .00 \), and literary and non-reading, F(1, 167) = .62, p = .43, \( \eta^2 = .00 \). The mean difference between the other conditions (i.e. non-fiction versus popular text, non-fiction versus non-fiction)
reading, popular fiction versus no reading) were all not significant ($p > .05$). The pairwise comparison results stayed almost the same when the analysis was repeated without dropping the interaction; literary versus non-fiction condition, $F(1, 151) = .06$, $p = .81$, $\eta_p^2 = .00$, literary versus popular fiction $F(1, 154) = .01$, $p = .93$, $\eta_p^2 = .00$, and literary versus non-reading condition, $F(1, 166) = .59$, $p = .44$, $\eta_p^2 = .00$.

A total of 51 participants reported their native language as other than English. When the data from these participants were excluded, the results remained similar to the previous results. We analysed RMET scores using repeated-measures ANOVA as before. The main effect for author recognition task was significant ($F(1, 268) = 10.47$, $p = .001$, $\eta_p^2 = .04$). The main effect of condition ($F(3, 268) = .27$, $p = .85$, $\eta_p^2 = .00$) and the interaction effect between the condition and the author recognition ($F(3, 268) = .56$, $p = .64$, $\eta_p^2 = .01$) were not significant.

We also conducted an analysis that controlled for participants’ experience with similar MTurk psychological experiments. When the data from participants with such experiences were excluded ($N = 92$), the results remained similar to the previous results. We analysed RMET scores using repeated-measures ANOVA as before. The main effect for author recognition task was significant ($F(1, 226) = 5.26$, $p < .05$, $\eta_p^2 = .02$). The main effect of condition ($F(3, 226) = .88$, $p = .45$, $\eta_p^2 = .01$) and the interaction effect between the condition and the author recognition ($F(3, 226) = 1.56$, $p = .20$, $\eta_p^2 = .02$) were not significant.

When we excluded participants on the criteria of both native language and previous MTurk participation at the same time, the results remained similar. We analysed RMET scores using repeated-measures ANOVA as before. The main effect for author recognition task was significant ($F(1, 187) = 8.54$, $p < .005$, $\eta_p^2 = .04$). The main effect of condition ($F(3, 187) = 1.01$, $p = .39$, $\eta_p^2 = .02$) and the interaction effect between the condition and the author recognition ($F(3, 187) = .82$, $p = .48$, $\eta_p^2 = .01$) were not significant. With this data, we also checked for condition by time interactions for the literary fiction condition in comparison to other conditions independently. There were no differences between literary fiction and non-fiction, $F(1, 111) = 1.40$, $p = .24$, $\eta_p^2 = .01$, literary and popular fiction, $F(1, 114) = 1.00$, $p = .32$, $\eta_p^2 = .01$, and literary and non-reading, $F(1, 126) = .14$, $p = .71$, $\eta_p^2 = .00$. The mean differences between the other conditions (i.e. non-fiction versus popular text, non-fiction versus no reading, popular fiction versus no reading) were all not significant ($p > .05$).

### Experiment 4

Experiment 4 was a pre-registered study as a direct replication of Experiment 3 by Kidd and Castano (2013). This experiment had twice as many participants (around 80 per condition), as the original study (around 35 per condition). We pre-registered this experiment because it is representative of the design and procedure for our previous studies as well as the studies by Kidd and Castano (2013). Because our primary interest was to understand if reading literary fiction enhances affective mentalising, we chose an experiment in which no cognitive task was applied. The original Experiment 3 by Kidd and Castano (2013) included a comparison between literary fiction and popular texts. We included two additional conditions, namely non-fiction and no-text, because the latter conditions were also used in the other experiments by Kidd and Castano (2013). Because we used a between subjects design, including these conditions did not compromise the integrity of the direct replication of the other conditions from the original article.

### Method

#### Participants and design

We recruited 321 participants for the final experiment through Mechanical Turk in the same way as in the previous experiments. In this experiment, we applied the same criterion with Experiment 3 for excluding participants based on low reading times (<30s/page). Based on this criterion, 22 participants were excluded. Applying the exact same outlier criteria from previous studies, two participants were excluded due to their low scores in RMET and three participants were excluded due to their high guessing rates in the author recognition test. The final data set consisted of 294 participants (161 Females; $M_{age} = 36.48$, range 18–68).

#### Materials

We included four conditions, in which participants read either literary fiction, popular fiction, non-fiction or no text. Every condition randomly used one of
the three texts described in the next lines. Stories in popular fiction condition included “Space Jockey” by Robert Heinlein (Hoppenstand, 1998), “Too Many Have Lived” by Dashiel Hammett (Hoppenstand, 1998) and “Lalla” by Rosamunde Pilcher (Hoppenstand, 1998). Stories in the literary fiction condition included “Corrie” by Alice Munroe (Furman, 2012), “Leak” by Sam Ruddick (Furman, 2012) and “Nothing Living Lives Alone” by Wendell Berry (Furman, 2012). The stories for the non-fiction condition were specifically chosen to match the word length of the other conditions. One of the texts was chosen from the original article, which is “How the Potato Changed the World” by Mann (2011). The other two texts were not included in the original article but they were taken from the same source. These texts were “Exploring the Titanic of the Ancient World” by Marchant (2015) and “Can the Siberian Tiger Make a Comeback” by Shaer (2015). Participants were randomly assigned to one of the readings listed above. We used the same materials as in Experiment 1 and 2.

**Procedure**

After reading the text, participants completed the RMET. Then participants completed the author recognition test and further demographic questions. At the end of the study, we included other personality questionnaires and the mood scale.

**Results**

The author recognition scores (M = 23.51, SD = 15.50) did not differ between conditions (F(3, 290) = 1.13, p = .34, ηp² = .01). The fantasy subscale from IRI was measured in all the conditions, and the scores did not differ between the conditions (F(3, 290) = .35, p = .79, ηp² = .00).

We analysed the RMET scores using ANOVA with condition as between participants factor and author recognition task as a covariate (see Table 1). As in previous experiments higher author recognition was associated with significantly better RMET performance r = .37, F(1, 289) = 47.45, p < .001, ηp² = .14. The main effect of condition remained non-significant F(3, 289) = 1.43, p = .24, ηp² = .02. The means and standard deviations for the conditions are shown in Table 2.

We also checked the effect for the literary fiction condition in comparison to other conditions independently. There were no differences between literary fiction and non-fiction, F(1, 128) = .47, p = .49, ηp² = .00, popular fiction, F(1, 125) = .76, p = .39, ηp² = .01, and non-reading, (1, 152) = .96, p = .33, ηp² = .01. The mean difference between other conditions (non-fiction versus popular text, non-fiction versus no reading, popular fiction versus no reading) were all not significant (p > .05). The pairwise comparison results stay almost the same when the analysis is done without dropping the interaction; literary versus non-fiction condition, F(1, 127) = .42, p = .52, ηp² = .00, literary versus popular fiction F(1, 124) = .78, p = .38, ηp² = .01, and literary versus non-reading condition, F(1, 151) = .96, p = .33, ηp² = .00.

We also checked for native language; four participants reported their native language as other than English. When this data was excluded, the results remained similar to the previous results. The main effect for author recognition task was significant (F(1, 282) = 43.86, p < .001, ηp² = .14). The main effect of condition (F(3, 282) = 1.38, p = .25, ηp² = .01) and the interaction effect between the condition and the author recognition (F(3, 282) = .09, p = .97, ηp² = .00) were not significant.

We also conducted an analysis that controlled for participants’ experience with the similar MTurk psychological experiments. When these data were excluded (N = 97), the results remained similar to the previous results. The main effect for author recognition was significant (F(1, 189) = 24.49, p < .001, ηp² = .12). The main effect of condition (F(3, 189) = .61, p = .61, ηp² = .01) and the interaction effect between the condition and the author recognition (F(3, 189) = .05, p = .98, ηp² = .00) were not significant.

When we applied the exclusion criteria of both native language and previous MTurk participation at the same time, the results from the same analysis remained similar. The main effect for author recognition task was significant (F(1, 186) = 23.57, p < .001, ηp² = .11). The main effect of condition (F(3, 186) = .67, p = .57, ηp² = .01) and the interaction effect between the condition and the author recognition (F(3, 186) = .05, p = .99, ηp² = .00) were not significant. With this data, we also checked for condition by time interactions for the literary fiction condition in comparison to other conditions independently. There were no differences between literary fiction and non-fiction, F(1, 107) = .04, p = .84, ηp² = .00, literary and popular fiction, F(1, 104) = 2.23, p = .14, ηp² = .02, and literary and non-reading, F(1, 131) = .02, p = .88, ηp² = .00. The mean differences
between the other conditions (i.e. non-fiction versus popular text, non-fiction versus no reading, popular fiction versus no reading) were all not significant ($p > .05$).

**Combined analyses**

Experiments 1–4 each failed to yield a statistically significant effect of reading literary fiction on mentalising performance. However, these null results should be interpreted against the statistical power of these experiments. We therefore performed a power analysis for the experiments from this article by using the G*Power 3.1.9.2 software (Faul, Erdfelder, Buchner, & Lang, 2009). The Cohen’s effect sizes ($f$) from the original Kidd and Castano (2013) experiments for Literary and Popular fiction comparison ranged from .15 to .26. Therefore, we calculated the power of our studies on the basis of both highest and lowest estimated effect sizes (see Appendix for details). Based on the highest effect size ($f = .26$), the power is estimated in between .83 and .92. Because a power of at least .80 is considered desirable, this means that our experiments had sufficient power to detect effects in the range of .26 and higher.

However, based on the lowest effect size ($f = .15$) of the original Kidd and Castano (2013), the power is estimated for all comparisons of literary fiction with other conditions in the range between .39 and .49. Thus, our experiments had insufficient statistical power to detect effects in the range of .15 and lower. Notably, this limitation also applied to the original experiments from Kidd and Castano (2013), which had a power in the range between .45 and .63 for the comparison of literary fiction with other conditions. The results of our post hoc power analysis thus indicates that lack of statistical power could pose a problem if the true effect size of reading literary fiction is very small.

Statistical power is a direct function of sample size. We therefore combined all the data together from Experiments 1–4 ($N = 1006$), so that our research would acquire the statistical power to detect even small effects. Because Experiment 3 was the only one to include a pre-manipulation measure of the RMET, we used the scores from the non-repeated items that were presented in post-manipulation for Experiment 3. To compare the different scores between studies, we further standardised the RMET scores by calculating the $z$-scores of the measure for each experiment. The standardised RMET scores were the main outcome for the combined analyses.

In the resulting data set, we had four conditions, namely literary fiction ($N = 296$), popular fiction ($N = 222$), non-fiction ($N = 302$) and no text ($N = 186$). Based on the lowest estimated effect size ($f = .15$), the power of all comparisons of literary fiction with other conditions is estimated to be at least .91. Based on the highest estimated effect size ($f = .26$), the estimated power reaches .99.

We analysed the RMET scores from the combined data using ANOVA with condition as between participants factor and author recognition test as a covariate. This analysis yielded a main effect for author recognition, $F(1, 998) = 254.31$, $p < .001$, $\eta_p^2 = .20$, no main effect of condition, $F(3, 998) = 1.03$, $p = .38$, $\eta_p^2 = .00$, and no interaction effect, $F(3, 998) = 1.63$, $p = .18$, $\eta_p^2 = .00$. When we ran the same analysis without Experiment 3, which was the only one to use a repeated RMET assessment, the results were highly similar.

We also conducted more focused tests of the effects of the literary fiction condition in comparison to other conditions independently in the combined data. There were no differences between literary fiction and non-fiction, $F(1, 594) = .24$, $p = .63$, $\eta_p^2 = .00$, popular fiction, $F(1, 514) = .04$, $p = .85$, $\eta_p^2 = .00$ and non-reading, $(1, 478) = 2.30$, $p = .13$, $\eta_p^2 = .01$. The mean differences between other conditions (non-fiction versus popular text, non-fiction versus no reading, popular fiction versus no reading) were all nonsignificant ($p > .05$). The pairwise comparison results stayed similar when the analysis was done for simple main effects; literary versus non-fiction condition, $F(1, 595) = .25$, $p = .62$, $\eta_p^2 = .00$, literary versus popular fiction $F(1, 515) = .03$, $p = .87$, $\eta_p^2 = .00$, and literary versus non-reading condition, $F(1, 479) = 2.41$, $p = .12$, $\eta_p^2 = .00$. Overall, the results of the combined analyses yielded the same pattern of nonsignificant effects for reading literary fiction.

To permit an additional check, we analysed the RMET scores from the combined data while applying the exclusion criteria on native speakers and naive MTurk participants from Study 3 and Study 4 ($N = 776$). Based on the lowest estimated effect size ($f = .15$), the power of all comparisons of literary fiction with other conditions was estimated to be in between .80 and .91. Based on the highest estimated effect size ($f = .26$), the estimated power reached .99. This analysis yielded a main effect for author
recognition, $F(1, 768) = 150.81, p < .001, \eta_p^2 = .16$, no main effect of condition, $F(3, 768) = .65, p = .58, \eta_p^2 = .00$, and no interaction effect, $F(3, 768) = 1.35, p = .26, \eta_p^2 = .01$. After dropping the interaction term, the simple main effects were similar for author recognition $F(1, 771) = 148.69, p < .001, \eta_p^2 = .16$ and condition $F(3, 771) = .83, p = .48, \eta_p^2 = .00$.

We also conducted more focused tests of the effects of the literary fiction condition in comparison to other conditions independently in the combined data as before. There were no differences between literary fiction and non-fiction, $F(1, 483) = .01, p = .92, \eta_p^2 = .00$, popular fiction, $F(1, 406) = .01, p = .93, \eta_p^2 = .00$, and non-reading, $(1, 347) = 1.43, p = .23, \eta_p^2 = .00$. The mean differences between other conditions (non-fiction versus popular text, non-fiction versus no reading, popular fiction versus no reading) were all nonsignificant ($p > .05$). The results of the combined analyses even with the strictest exclusion criteria yielded the same pattern of nonsignificant effects.

**General discussion**

Does a single session of reading literary fiction lead to better mentalising performance? To address this question, we conducted four close replications of Kidd and Castano (2013). In these experiments, we manipulated whether participants read literary fiction versus non-fiction (Experiments 1–4), popular fiction (Experiments 2–4), or no text (Experiments 3 and 4). After this, participants completed the RMET, a well-established measure of mentalising performance. In each of these four experiments, we found no significant effects of reading literary fiction on mentalising performance. Moreover, analysing the combined data of four experiments with a total of $N = 1017$ participants yielded no significant effects of reading literary fiction. The present results thus fail to confirm the notion that a single session of reading literary fiction can prime enhanced mentalising performance.

Our findings in the present experiments differ markedly from Kidd and Castano (2013), who reported statistically significant priming effects of reading literary fiction in five experiments (combined $N = 697$). What would be the reasons for these discrepant findings? First, the present research was sufficiently powered to detect small effects, particularly in the analysis that combined the data of all our 1006 participants. Thus, the present null findings are not due to lack of statistical power. Second, the present experiments used the same set of texts in literary and popular conditions and mostly the same texts in non-fiction conditions, same procedures, and same Amazon Mechanical Turk samples as the original experiments. It therefore seems unlikely that subtle differences in procedure, stimulus materials, or samples can explain the different findings in the present studies and in Kidd and Castano. Third, we replicated the correlation between familiarity with literature and mentalising performance that was also observed by Kidd and Castano (2013) and in prior research (Mar et al., 2009; Mar & Oatley, 2008). This finding indicates that our procedures were generally valid and sensitive enough to detect the relation between reading fiction and mentalising performance.

As mentioned earlier, while we were in the process of publishing our results, another replication attempt of the same article by Kidd and Castano (2013) has been published by Panero and her colleagues (2016). In this article, they recruited 792 participants and assigned them randomly to four conditions (literary fiction, popular fiction, non-fiction, and no reading), as it is in the original and the present article. They were also unable to find higher mentalising scores after reading literary fiction compared to other conditions. Additionally, they were also able to replicate the correlation between RMET scores and Author Recognition test, as it is in the present article. This further evidence strongly converges with the present findings.

A closer examination of our results and the results of Kidd and Castano (2013) indicates that mean RMET scores in the present studies ($M = 24.97, \text{range} = 23.22–26.21$) were several points lower than in the Kidd and Castano studies ($M = 24.97, \text{range} = 23.22–26.21$). This apparent discrepancy could signify a systematic difference between the two sets of studies. This discrepancy was also evident in the other previously mentioned replication article by Panero and her colleagues (2016). Moreover, they also reported higher overall RMET scores, similar to ours. In our Experiment 3, repeated testing of the RMET appeared to lower RMET scores, suggesting that RMET performance may be susceptible to internal motivational influences, such as fatigue. Moreover, the willingness of people to effortfully engage in mentalising is not the same as the mind-reading ability (Carpenter, Green, & Vacharkulksemsuk, 2016), and it is shown that mind-reading motivation can be influenced from an external information, such as the source of the text (selected by researchers vs. computer-generated).
Even though the original authors focused on a priming account, it is conceivable that literary fiction might increase internal and external motivation to use their mentalising abilities. Even though we can't point to a specific circumstance to explain the difference between our mentalising scores and the original article, future work should consider the potential role of motivation in the effects of reading literary fiction.

One important limitation of the present studies is that they used only a limited selection of literary, popular and non-fiction texts. Because the present studies were designed to closely replicate the Kidd and Castano (2013) paradigm, we did not systematically consider the substantive contents of the texts. However, it is possible that the results would be different if a different selection of texts were considered. Furthermore, in future studies, qualitative and quantitative analyses should be performed on the experimental materials in order to show what it is in the texts that requires more mentalising effort and what not.

Furthermore, we included a minor methodological divergence in Experiment 3 with introducing a pre-RMET measure. Given that the original Kidd and Castano studies did not include a baseline measure, their findings could have also been due to a lack of control for baseline differences. Indeed, the results in Experiment 3 show that such baseline differences can sometimes erroneously suggest that there is a significant effect of reading fiction on mentalising performance. Such potential artefacts should be carefully controlled in future studies. Another limitation is that we did not assess previous exposure to the readings. This variable should be included in future studies.

Finally, it is important to recognise that the present findings are limited to the effects of a single session of reading literary fiction, in line with the priming paradigm developed by Kidd and Castano (2013). The Kidd and Castano paradigm, and priming effects more generally, are rather minimalistic, given that they focus only on the effects of a one shot exposure to the priming stimulus, in the present case, reading literary fiction. It will be important to incorporate richer paradigms that are conducted over longer periods of time in future research on the psychological effects of reading literary fiction. Indeed, some prior work suggests that the psychological effects of reading fiction may increase over time, (Appel & Richter, 2007; Bal & Veltkamp, 2013), perhaps because these effects have to be consolidated in memory over time (Schank & Abelson, 1995). It is possible that more enriched and prolonged exposures to literary fiction bring greater psychological benefits.

Note
1. The full texts for all the stories/excerpts used in all experiments are available by request. Please contact the correspondence author for a copy.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding information
The authors gratefully acknowledge the support of a Consolidator Grant from the European Research Council [ERC-2011-StG_20101124] to Sander L. Koole and a Research Talent Grant from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek [NWO-406-14-097] to Dalya Samur and Sander L. Koole.

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(1), 278–289.
Adler, J., & Lawler, A. (2012, June). How the Chicken Conquered the World. Smithsonian Magazine.
Appel, M., & Richter, T. (2007). Persuasive effects of fictional narratives increase over time. Media Psychology, 10(1), 113–134.
Bal, P. M., & Veltkamp, M. (2013). How does fiction reading influence empathy? An experimental investigation on the role of emotional transportation. PloSone, 8(1), e55341.
Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. Journal of child psychology and psychiatry, 42(2), 241–251.
Baumann, N., Kaschel, R., & Kuhl, J. (2005). Striving for unwanted goals: Stress-dependent discrepancies between explicit and implicit achievement motives reduce subjective well-being and increase psychosomatic symptoms. Journal of personality and social psychology, 89(5), 781–799.
Black, J. E., & Barnes, J. L. (2015). The effects of reading material on social and non-social cognition. Poetics, 52, 32–43.
Bormann, D., & Greitemeyer, T. (2015). Immersed in virtual worlds and minds effects of in-game storytelling on immersion, need satisfaction, and affective theory of mind. Social Psychological and Personality Science, 1948550615578177.
Carpenter, J. M., Green, M. C., & Vacharkulksemsuk, T. (2016). Beyond perspective-taking: Mind-reading motivation. Motivation and Emotion, 40(3), 358–374.
Chekhov, A. (1979). Antion Chekov’s Short Stories. 3.
Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple correlation/regression analysis for the behavioral sciences. Mahwah, NJ: Lawrence Erlbaum Associates.
D. Samur et al. (2011). The collected stories of Lydia Davis. New York: Penguin UK.

Davis, M. H., & Association, A. P. (1980). A multidimensional approach to individual differences in empathy.

De Dreu, C. K., Koole, S. L., & Steinel, W. (2000). Unfixing the fixed pie: A motivated information-processing approach to integrative negotiation. Journal of personality and social psychology, 79(6), 975–987.

De Lillo, D. (2011). The Angel Esmeralda: Nine stories. New York: Simon and Schuster.

Djikic, M., Oatley, K., & Moldoveanu, M. C. (2013). Reading other minds: Effects of literature on empathy. Scientific Study of Literature, 3(1), 28–47.

Dunn, R. (2011). Mistletoe: The evolution of a Christmas tradition. Smithsonian.com. Retrieved from http://www.smithsonian.com/science-nature/mistletoe-the-evolution-of-a-christmas-tradition-10814188/.

Dunn, R. (2012, March). The story of the most common bird in the world. Smithsonian Magazine. Retrieved from http://www.smithsonianmag.com/science-nature/the-story-of-the-most-common-bird-in-the-world-113046500/.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160.

Flynn, G., Whelan, J., & Heyborne, K. (2012). Gone Girl: Phoenix.

Forman, L. (2012). The Pen/O. Henry Prize stories 2012: The best stories of the year. New York: Anchor.

Gandel, C. (2008, March). Bamboo steps up. Smithsonian Magazine. Retrieved from http://www.smithsonianmag.com/greener-living/bamboo-steps-up-33166568/.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. Journal of Personality and Social Psychology, 79(5), 701–721.

Hakemulder, J. (2000). The moral laboratory: Experiments examining the effects of reading literature on social perception and moral self-concept (Vol. 34). Amsterdam: John Benjamins.

Hoppenstand, G. (1998). Popular fiction: an anthology. New York: Longman.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. Science, 342(6156), 377–380.

Koole, S. L., & Lakens, D. (2012). Rewarding replications a sure and simple way to improve psychological science. Perspectives on psychological science, 7(6), 608–614.

Kuiken, D., Miall, D. S., & Sikora, S. (2004). Forms of self-implication in literary reading. Poetics Today, 25(2), 171–203.

Mann, C. C. (2011, November). How the potato changed the world. Smithsonian Magazine. Retrieved from http://www.smithsonianmag.com/history/how-the-potato-changed-the-world-108470605/.

Marchant, J. (2015, February). Exploring the Titanic of the ancient world. Smithsonian Magazine.

Obrecht, T. (2011). The tiger’s wife: A novel. New York: Random House Incorporated.

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(5), 446–e54.

Roediger, H. (2012). Psychology’s woes and a partial cure: The value of replication. APS Observer, 25(2), 9, 27–29.

Samur, D., Tops, M., Schlinkert, C., Quirin, M., Cuijpers, P., & Koole, S. L. (2013). Four decades of research on alexithymia: Moving toward clinical applications. Frontiers in Psychology, 4, 861, 1–4.

Schank, R. C., & Abelson, R. P. (1995). Knowledge and memory: The real story. Knowledge and Memory: The Real Story. Advances in Social Cognition, 8, 1–85.

Shaer, M. (2015, February). Can the Siberian tiger make a comeback? Smithsonian Magazine.

Smith, M. C. (1990). Reading habits and attitudes of adults at different levels of education and occupation. Literacy Research and Instruction, 30(1), 50–58.

Tucker, A. (2010, February). The Venus flytrap’s lethal allure. Smithsonian Magazine.

Vorst alexithymia questionnaire.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., Kievit, R. A. (2012). An agenda for purely confirmatory research. Perspectives on Psychological Science, 7(6), 632–638.

Appendix A. Power Analysis

We conducted a post hoc power analysis by using the software G*Power 3.19.2 (Faul et al., 2009). To perform the analysis, first we converted the effect sizes to Cohen’s f, which was required by the software we used.

We performed the power analysis which aimed to compute achieved power based on the given alpha, sample size and interactions based on the statistical test ANCOVA (Fixed effects, main effects, and interactions). We calculated the power for every pairwise comparison with literary fiction and other conditions for all experiments that included RMET as the dependent variable (see Table 1).

To perform the pairwise comparisons, we needed the exact sample size for all conditions. Since this was not reported in Experiment 5 from the original article, we calculated an estimation per condition under the assumption that all conditions had equal number of participants.
We further calculated the power analysis from the present article. The Cohen’s effect sizes ($f$) from the original Kidd and Castano (2013) experiments for literary and popular fiction comparison ranged from .15 to .26. Therefore, we calculated the power of our studies on the basis of both lowest and highest estimated effect sizes.

**Table A2.** Power analysis for the experiments from the present article.

| Condition 1 | Condition 2 | Sample size | Power ($f = .15$) | Power ($f = .26$) |
|-------------|-------------|-------------|-------------------|-------------------|
| Experiment 1 Literary fiction | Non-fiction | 156 | .46 | .90 |
| Experiment 2 Literary fiction | Non-fiction | 156 | .46 | .90 |
| Experiment 2 Literary fiction | Popular fiction | 154 | .46 | .89 |
| Experiment 3 Literary fiction | Non-fiction | 155 | .46 | .90 |
| Experiment 3 Literary fiction | Popular fiction | 158 | .47 | .90 |
| Experiment 3 Literary fiction | No-reading | 170 | .49 | .92 |
| Experiment 4 Literary fiction | Non-fiction | 131 | .46 | .90 |
| Experiment 4 Literary fiction | Popular fiction | 128 | .39 | .83 |
| Combined data Literary fiction | Non-fiction | 598 | .96 | .99 |
| Combined data Literary fiction | Popular fiction | 518 | .93 | .99 |
| Combined data Literary fiction | No-reading | 482 | .91 | .99 |
| Combined data * Literary fiction | Non-fiction | 487 | .91 | .99 |
| Combined data * Literary fiction | Popular fiction | 410 | .86 | .99 |
| Combined data * Literary fiction | No-reading | 351 | .80 | .99 |

*Combined data with exclusion criteria on native speakers and naive participants applied for Experiment 3 and Experiment 4

### Appendix B. List of Readings from the Present article and the Kidd and Castano (2013) article

**Table B1.** The list of Readings from the Present article and the original article, Kidd and Castano (2013).

| Present article | Literary fiction | Popular fiction | Non-fiction |
|----------------|------------------|----------------|------------|
| Experiment 1   | “The Runner” by Don DeLillo | x | “How the Potato Changed the World” by Charles C. Mann |
|                | “Blind Date” by Lydia Davis | x | “Bamboo Steps Up” by Cathie Gandel |
|                | “Chameleon” by Anton Chekhov | x | “The Story of the Most Common Bird in the World” by Rob Dunn |
| Experiment 2   | “The Tiger’s Wife” by Téa Obreht | “Gone Girl” by Gillian Flynn | “How the Chicken Conquered the World” by Jerry Adler and Andrew Lawler * |
|                | “Uncle Rock” by Dagoberto Gilb | “Space Jockey” by Robert Heinlein | “Mistletoe: The Evolution of a Christmas Tradition” by Rob Dunn * |
|                | “The Vandercook” by Alice Mattinson | “Jane” by Mary Roberts Rinehart | “The Venus Flytrap’s Lethal Allure” by Abigail Tucker * |
| Experiment 3   | “The Runner” by Don DeLillo | “Gone Girl” by Gillian Flynn | “How the Potato Changed the World” by Charles C. Mann |
|                | “Uncle Rock” by Dagoberto Gilb | “Space Jockey” by Robert Heinlein | “The Story of the Most Common Bird in the World” by Rob Dunn |
|                | “The Vandercook” by Alice Mattinson | “Jane” by Mary Roberts Rinehart | “How the Chicken Conquered the World” by Jerry Adler and Andrew Lawler * |
| Experiment 4   | “Corrie” by Alice Munro | “Too Many Have Lived” by Dashiel Hammett | “How the Potato Changed the World” by Charles C. Mann |
|                | “Leak” by Sam Ruddick | “Space Jockey” by Robert Heinlein | “Exploring the Titanic of the Ancient World” by Jo Marchant * |
|                | “Nothing Living Lives Alone” by Wendell Berry | “Lalla” by Rosamunde Pilcher | “Can the Siberian Tiger Make a Comeback” by Matthew Shaer * |

(Continued)
| Present article | Literary fiction | Popular fiction | Non-fiction |
|-----------------|-----------------|-----------------|------------|
| **Ridd and Castano (2013)** | | | |
| Experiment 1 | “The Runner” by Don DeLillo | x | “How the Potato Changed the World” by Charles C. Mann |
| | “Blind Date” by Lydia Davis | x | “Bamboo Steps Up” by Cathie Gandel |
| | “Chameleon” by Anton Chekhov | x | “The Story of the Most Common Bird in the World” by Rob Dunn |
| Experiment 2 | “The Tiger’s Wife” by Téa Obreht | | “Gone Girl” by Gillian Flynn |
| | “The Round House” by Louise Erdrich | | “The Sins of the Mother” by Danielle Steel |
| | “Salvage the Bones” by Jesmyn Ward | | “Cross Roads” by W. Paul Young |
| Experiment 3 | “Corrie” by Alice Munro | | “Too Many Have Lived” by Dashiell Hammett |
| | “Leak” by Sam Ruddick | | “Space Jockey” by Robert Heinlein |
| | “Nothing Living Lives Alone” by Wendell Berry | | “Lalla” by Rosamunde Pilcher |
| Experiment 4 | “Uncle Rock” by Dagoberto Gilb | | “Jane” by Mary Roberts Rinehart |
| | “Corrie” by Alice Munro | | “Too Many Have Lived” by Dashiell Hammett |
| | “Nothing Living Lives Alone” by Wendell Berry | | “Space Jockey” by Robert Heinlein |
| Experiment 5 | “The Vandercook” by Alice Mattinson | | “Jane” by Mary Roberts Rinehart |
| | “Uncle Rock” by Dagoberto Gilb | | “Too Many Have Lived” by Dashiell Hammett |
| | “Corrie” by Alice Munro | | “Space Jockey” by Robert Heinlein |

**Table B1.**