How Fiction Disrupts

David Kidd and Emanuele Castano

Harvard Graduate School of Education
and
Dipartimento di Sciencze Umane e Sociali, Universita degli Sudi de Bergamo (Italy)

This project was supported in part or in whole by an award from the *Research: Art Works* program at the National Endowment for the Arts: Grant# 16-3800-7004.

Article

Reading Literary Fiction and Theory of Mind: Three Preregistered Replications and Extensions of Kidd and Castano (2013)

Social Psychological and Personality Science
1-10

DOI: 10.1177/1948550618775410
journals.sagepub.com/home/spp
SAGE

David Kidd[1] and Emanuele Castano[2]

Abstract

Scholars from diverse disciplines have proposed that reading fiction improves intersubjective capacities. Experiments have yielded mixed evidence that reading literary fiction improves performance on the Reading the Mind in the Eyes Test, a test of Theory of Mind. Three preregistered experiments revealed mixed results. Applying the "small telescopes" method developed by Simonsohn revealed two uninformative failures to replicate and one successful replication. On a measure of the importance of intentions to moral judgments, results were more mixed, with one significant effect in the expected direction, one nonsignificant effect, and one significant effect in the unexpected direction. In addition, two experiments yielded support for the exploratory but preregistered hypothesis that characters in popular fiction are perceived as more predictable and stereotypic than those in literary fiction. These findings help clarify the sociocognitive effects of reading literary fiction and refine questions for future research.

Keywords

morality, reading, social cognition, sociocultural factors, theory of mind

> When I think about how I understand my role as citizen, setting aside being president, and the most important set of understandings that I bring to that position of citizen, the most important stuff I've learned I think I've learned from novels. It has to do with empathy. It has to do with being comfortable with the notion that the world is complicated and full of grays, but there's still truth there to be found, and that you have to strive for that and work for that. (Obama & Robinson, 2015)

Barack Obama's speculation that reading fiction has honed his appreciation for the nuances of others' experiences, and the effort needed to discern them is not idiosyncratic. Thinkers ranging from the cognitive developmental psychologist Jerome Bruner (1986) to the philosopher Martha Nussbaum (1985, 1991) posit that fiction prompts its readers to challenge presuppositions, engage with other perspectives, and check the tendency to form quick and simple verdicts. These would be notable achievements. As we navigate our social world, stereotypes and norms efficiently reduce its overwhelming complexity (Hirschfeld, 2006), and we fully deploy the more cognitively costly process of engaging with others' minds, or Theory of Mind (ToM), only when the more schematic strategies prove insufficient or we are motivated to establish or maintain a relationship (Fiebich & Coltheart, 2015).

Whether reading fiction shifts our habitual reliance on general social knowledge, or *theory of society* (Hirschfeld, 2006), to a mode of social perception that emphasizes ToM may depend on how authors depict social content (Kidd & Castano, 2013; Kidd & Castano, 2017a; Kidd, Ongis, & Castano, 2016). Encountering stereotypical (or stock) characters and formulaic situations, readers can draw on well-rehearsed social stereotypes or genre-specific expectations (Culpeper, 2001; Schneider, 2001). By contrast, more subtly represented and nuanced relationships or characters require that readers carefully attend to cues to characters' mental states, interpreting and reinterpreting them as more information is made available (Zunshine, 2015a, 2015b). As literary scholars have observed, such complexity is most readily found in literary fiction (Eder, Jannidis, & Schneider, 2010; Hakemulder, 2000; Miesen, 2004). In contrast, popular genre fiction features more formulaic characters and social situations, and the reader's attention is usually focused on plot development or vicarious emotions resulting from strong identification with relatively straightforward characters (Gelder, 2004; Keen, 2011). Ordinary readers

[1] Harvard Graduate School of Education, Harvard University, Cambridge, MA, USA
[2] Dipartimento di Scienze Umane e Sociali, Università degli Studi di Bergamo, Bergamo, Italy

Corresponding Author:
David Kidd, Project Zero, Harvard School Graduate of Education, Harvard University, 13 Appian Way, Cambridge, MA 02138, USA.
Email: david_kidd@gse.harvard.edu

appear to recognize this, seeking more literary works for experiences of insight, and genre fiction for entertainment (Miesen, 2004). Some evidence suggests that readers may also adopt different modes of engagement based on cues to the text's genre (Gavaler & Johnson, 2018).

Drawing on this distinction, Kidd and Castano (2013) tested the hypothesis that reading literary fiction would lead to improved ToM performance in five experiments, four of which directly contrasted literary and popular fiction. Consistent with the hypothesis, results revealed higher ToM performance among participants assigned to read literary fiction (see also Black & Barnes, 2015; Kidd et al., 2016; Pino & Mazza, 2016; van Kuijk, Verkoeijen, Dijkstra, & Zwaan, 2018). A subsequent series of large-scale correlational studies further demonstrated that the reliable finding that lifetime exposure to fiction positively predicts ToM performance (Mar, Oatley, Hirsh, dela Paz, & Peterson, 2006; Mar, Oatley, & Peterson, 2009; Mumper & Gerrig, 2017) is driven by exposure to literary fiction (Kidd & Castano, 2017a). Some studies have failed to demonstrate the experimental effect of reading literary fiction on ToM (Panero et al., 2016; Samur, Koole, & Topps, 2018), though a reanalysis of one of the studies suggests that the null finding may be in part due to methodological differences from the original studies (Kidd & Castano, 2017b). Thus, experiments designed to precisely replicate the relative impact of reading literary fiction compared to popular fiction on ToM are critical.

The extant correlational and experimental research has operationalized ToM performance mostly with the Reading the Mind in the Eyes Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), an advanced test of ToM that, though extensively used and well validated (Vellante et al., 2013), has been criticized due to its positive relation with verbal intelligence (Peterson & Miller, 2012). Furthermore, it appears that only one experiment (Pino & Mazza, 2016) and one correlational study (Tamir, Bricker, Dodell-Feder, & Mitchell, 2016) utilized a ToM task that did not involve detection of mental states from images of faces or people. There is therefore a need to test the generality of the effects of reading literary fiction on ToM. Thus, a moral judgment paradigm that is thought to evoke ToM was also included to test whether effects could be observed in this context.

Although the main purpose of these studies is to test the effect of reading literary fiction on the RMET, they also provide an opportunity to test a somewhat contentious theoretical assumption: That characters in literary fiction are less easily understood in schematic terms than those in popular genre fiction. Although grounded in literary studies and empirical studies of readers' preferences as described above, some scholars have argued that the distinction between literary and popular fiction is unclear (Gavaler & Johnson, 2018) and questioned whether readers experience the types of fiction in consistently different ways (Panero et al., 2016). The second and third experiments directly test the hypothesis that readers perceive literary characters as less predictable, stereotypic, and easily understood than characters in popular genre fiction.

The present three studies were designed primarily to test the effect of reading literary as compared to popular genre fiction on RMET performance. To directly assess consistency with the results of Experiment 5 in Kidd and Castano (2013), a "small telescopes" analysis was conducted to compare replication effect sizes with those in the original study (Simonsohn, 2015). At the request of the editor, this approach was adopted in place of the internal meta-analysis mentioned in the third registration. Secondary aims include testing the effect of reading literary fiction on a measure of the importance of inferred intentions to moral judgments and, in two experiments, testing the hypothesis that readers perceive characters in popular genre fiction as more clearly defined than those in literary fiction.

Method

These experiments are preregistered replications and extensions of the fifth experiment reported in Kidd and Castano (2013). The experiments also included a moral judgment paradigm that is thought to involve ToM processes (Young, Cushman, Hauser, & Saxe, 2007). In Experiments 2 and 3, a 7-item exploratory measure of character perception was also administered. Copies of the preregistrations, study materials, data, and analysis syntax are available via hyperlink in the Online Supplemental Materials and on the Open Science Foundation server (https://osf.io/yvud8/).

Materials and Procedure

Experiment 1

Procedure. Participants were randomly assigned to one of the three conditions: literary fiction, popular fiction, and baseline. Participants in the two reading conditions were randomly assigned to read one of the three texts associated with that condition, and participants in the baseline condition read nothing. Participants then completed the RMET (Baron-Cohen et al., 2001), the moral judgment task (MJT; Moran et al., 2011; Young et al., 2007), the Author Recognition Test (ART; Acheson, Wells, & MacDonald, 2008), and a brief demographics questionnaire. The demographics questionnaire included the following statement:

> This is one of several studies we have conducted on this topic. If you think that you have participated in a very similar study, please let us know. It will ***not*** affect your compensation if you have participated in the past, so please be honest—you will still receive full compensation.

Then participants were given three response options: "Yes, I think I have participated in a similar study;" "No, I do not think I have participated in a similar study;" and "I have heard about this research, but I have not participated in any studies similar to this one." Only participants who selected, "No, I do not think I have participated in a similar study," were considered naive. All participants were debriefed, thanked, and compensated.

Short stories. The same six texts were used as in Experiment 5 of Kidd and Castano (2013). The literary fiction condition

included short stories selected from winners of the 2012 PEN/O. Henry Prize (Furman, 2012): *Corrie* by Alice Munro,, *The Vandercook* by Alice Mattison, and *Uncle Rock* by Dagoberto Gilb. Short stories in the popular genre fiction condition were selected to represent the different genres from an edited anthology of popular fiction (Hoppenstand, 1998): *Jane* by Mary Roberts Rinehart (romance), *Space Jockey* by Robert Heinlein (science fiction), and *Too Many Have Lived* by Dashiell Hammett (thriller/mystery).

RMET. The RMET (Baron-Cohen et al., 2001) contains 36 trials in which participants select which of the four emotion terms best matches an image of an actor's eyes. Scores reflect the number of correct matches. This measure has been used in prior correlational (e.g., Kidd & Castano, 2017a; Mar et al., 2009) and experimental studies (e.g., Black & Barnes, 2015; Kidd & Castano, 2013; Panero et al., 2016) of the relation between reading fiction and ToM. Critically, it is the ToM measure used in Experiment 5 of Kidd and Castano (2013). Although definitions for RMET terms are often provided to participants, they were not in Kidd and Castano (2013) or in this replication.

MJT. This task directs participants to rate the moral permissibility of actions that varies according to their outcome (neutral vs. negative) and the actor's beliefs about the likely outcome (neutral vs. negative). As Moran et al. (2011) point out, normal adults typically rate accidental harms (neutral belief + negative outcome) as more permissible than attempted harms (negative belief + neutral outcome). The extent to which they do so can be considered an index of the extent to which they infer and prioritize actors' intentions when making moral judgments (Young et al., 2007; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010; Young & Saxe, 2008), and responses to the task have been shown to relate to familiarity with fiction (Tamir et al., 2016). To avoid participant fatigue, an abbreviated version of the task was used, with three scenarios representing each of the four conditions, rather than the usual 12 (e.g., Tamir et al., 2016; Young et al., 2010). ToM scores on this task were calculated as indicated in the preregistration by subtracting moral permissibility ratings for failed harms from those of accidental harms.

ART. The ART (Acheson et al., 2008) is a checklist containing 130 names, half of which are those of fiction authors and half of which are foils. Participants are instructed to "please put a check mark next to the [names] that you know for sure are authors" (for full instructions, see https://osf.io/yvud8/). Scores are calculated by subtracting the number of foils selected from the number of authors selected. ART measures have been widely used in reading research, and they correlate well with actual reading habits (Rain & Mar, 2014).

Experiment 2. The materials and procedure for the second experiment were identical to those in the first experiment except for two differences. First, the baseline condition was excluded in this experiment to allow for a more powerful comparison of the literary and popular fiction conditions. Second, a 7-item measure of readers' perceptions of characters, the Character Clarity Scale (CCS), was administered after the MJT and before the ART. Four of the 7 items in the CCS were developed to assess the perceived typicality, predictability, and ambiguity of the character. Two items assessed participants' confidence in their personality ratings of characters. Each item was responded to on a sliding scale of 1–7, and responses were averaged after 3 items were reverse coded (see Online Supplemental Materials for all items).

Factor analyses in both experiments including the scale revealed two eigenvalues greater than 1, suggesting two factors. In both experiments, an item directly assessing perceived character complexity was the only item that loaded more strongly onto the second factor than the first, and the only item with a loading of less than .4 on the first factor. Consequently, this item was dropped from the final CCS, yielding acceptable Cronbach's αs in Experiments 2 ($\alpha = .72$) and 3 ($\alpha = .74$).

Experiment 3. The procedures and materials in Experiment 3 were identical to those in Experiment 2, except for the addition of additional questionnaires included for exploratory purposes and not presented here (see preregistration).

Participants

Experiment 1. Participants were recruited and compensated using Qualtrics Panels, a third-party data collection service. Early in the data collection process, Qualtrics sent a preliminary data set for review. This revealed a large imbalance among the conditions, with nearly twice as many baseline participants as participants in either of the reading conditions. Qualtrics was asked to stop collecting baseline data once 100 responses were obtained. This cutoff was chosen to ensure sufficient participants to test the key hypothesis regarding the difference between the reading conditions. Although contracted to obtain 450 responses, Qualtrics delivered a data set including 479 responses (478 were complete).

The preregistered inclusion criteria were applied to yield the final sample (see Table 1). All distribution-based exclusions were made using means and standard deviations calculated after applying all prior exclusions. Unexpectedly, some participants indicated on the MJT that actions with neutral outcomes motivated by neutral intentions were less morally permissible than harmful actions motivated by harmful intent. No exclusion criterion to remove these abnormal responses was preregistered for this first study, but it was added to the subsequent registrations. To maintain consistency, results for the first experiment are presented first as preregistered and then with this exclusion criterion applied.

A power analysis of the final sample ($N = 305$) conducted using G*Power (Faul, Erdfeld, Lang, & Buchner, 2007) indicated 80% power to detect the effect observed in the original study ($d = .33$). However, the distribution of participants across conditions was not even. While the literary fiction ($n = 137$) and popular fiction conditions ($n = 110$) were roughly equal in size, the baseline condition included only 58 participants. This unintended imbalance threatens the internal validity of

Table 1. Participant Exclusions.

Exclusion Criterion	Experiment 1 ($N = 478$)	Experiment 2 ($N = 381$)	Experiment 3 ($N = 432$)
Prior/nonnaive/ undisclosed participation	103 21.55%	25 6.56%	37 8.56%
First language other than English	13 2.72%	2 0.52%	17 3.94%
ART scores ≤ 0	21 4.39%	12 3.15%	6 1.39%
RMET < 9	1 0.21%	0 0.00%	0 0.00%
Reading time < 30s/ page	20 4.18%	13 3.41%	19 4.40%
Neutral scenarios rated less acceptable than intentionally harmful scenarios	5 1.05%	3 0.79%	1 0.23%
ART guesses ≥ 3.5 *SD* from mean	7 1.46%	6 1.57%	4 0.93%
RMET ≥ 3.5 *SD* from mean	0 0.00%	1 0.26%	0 0.00%
Reading times ≥ 3.5 *SD* from mean	3 0.63%	1 0.26%	2 0.46%
Total exclusions	173 36.19%	63 16.54%	86 19.91%
Final sample	305	318	346

Note. ART = Author Recognition Test; RMET = Reading the Mind in the Eyes Test; *SD* = standard deviation.

comparisons with the baseline condition. Therefore, these comparisons are reported only in the Online Supplemental Materials. Participant characteristics are presented in Table 2.

Experiment 2. Prolific Academic (http://www.Prolific.ca) was contracted to recruit 400 participants who were compensated $5.75 for their participation. On Qualtrics, 382 responses were recorded (381 complete). After applying the preregistered exclusion criteria (see Table 1), a final sample of 318 was retained (see Table 2 for descriptive statistics). This sample had 83% power to detect the effect observed in the original study ($d = .33$).

Experiment 3. In the third experiment, 400 participants were recruited using Amazon Mechanical Turk (http://www.Mturk.com), and 448 responses were recorded (432 complete). After applying the preregistered exclusion criteria (see Table 1), 346 participants were retained (see Table 2 for participant characteristics), yielding 86% power to detect the effect observed in the original study ($d = .33$).

Results

RMET

In each experiment, RMET scores were entered as the dependent variable in a Generalized Linear Model (GLM) with experimental condition (literary fiction, popular genre fiction, and baseline) and ART scores (square-root transformed and standardized) entered as full factors (see Table 3). The main effect of condition was qualified by an interaction with ART in Experiment 1, not significant in Experiment 2, and significant in Experiment 3 (see Table 4 for least squares means). Following the request of the editor, the meta-analysis of the combined data (included in the registration for Experiment 3) was not conducted. Instead, the effect of primary concern, that of reading condition on RMET performance, from each study was compared to that observed in the original (Experiment 5 in Kidd & Castano, 2013) using the detectability, or "small telescopes," test developed by Simonsohn (2015). This test compares the confidence interval (CI)of a replication effect size (i.e., Cohen's *d*) to the effect size that could have been detected in the original study with 33% power ($d_{33\%}$). This test informs the evaluation of replication effects by showing whether a significant replication effect is too small to have been observed in the original study. It also indicates whether a nonsignificant replication effect is evidence in favor of the null hypothesis or is merely inconclusive. In the case of a significant replication effect, an effect size with a CI not including the original effect size is interpreted as not replicating the original finding, since it would have low probability of being detected in the original study. Nonetheless, the effect may be true, even if the original study could not have detected it due to low statistical power. When a replication effect is not significant but the CI of the effect includes effects larger than $d_{33\%}$, the replication is inconclusive. If the CI of the replication effect includes zero but not $d_{33\%}$, it demonstrates both a clear failure to replicate the original effect and evidence in favor of the null hypothesis.

To calculate $d_{33\%}$, a sensitivity analysis was conducted using G*Power (Faul et al., 2007) using the sample size ($N = 225$) from Experiment 5 of Kidd and Castano (2013), power of 33%, and a two-tailed test ($\alpha = .05$). This yielded $d_{33\%} = .203$. For each of the three effects of reading condition on RMET performance observed in the present studies, 95% CIs of the effect sizes (see Table 5) were computed using the SAS syntax provided in the supplementary materials of Simonsohn (2015). Since the CIs of the effect sizes from the two nonsignificant replication attempts include $d_{33\%}$, they are inconclusive or uninformative. The third study is a successful replication, with the observed effect size and the upper bound 95% CI both above $d_{33\%}$.

MJT

In the first experiment, scores on the MJT were entered as the dependent variable in a GLM with experimental condition and ART scores entered as full factor independent variables. The MJT also includes two conditions in which intentions and outcomes are congruent, and so a baseline moral judgment score was calculated by subtracting the moral permissibility ratings of intended harms from those for neutral scenarios. Although not preregistered for the first experiment, including this baseline judgment score as a covariate seems to provide a clearer test of the hypothesis, and it was included in the registrations

Table 2. Variable Means and Demographic Information.

	M (*SD*) 95% CIs		
Characteristic	Experiment 1	Experiment 2	Experiment 3
RMET	25.49 (4.16) [25.02, 25.96]	26.61 (4.67) [26.09, 27.12]	26.57 (3.93) [26.15, 26.99]
Moral mind	1.34 (1.63) [1.16, 1.53]	1.70 (1.62) [1.52, 1.88]	1.51 (1.67) [1.34, 1.69]
Moral base	3.30 (0.93) [3.19, 3.40]	3.49 (0.74) [3.41, 3.57]	3.45 (0.68) [3.37, 3.52]
ART	23.11 (15.39) [21.38, 24.84]	20.95 (13.85) [19.42, 22.48]	17.89 (12.73) [16.54, 19.23]
Age	55.61 (16.55) [53.75, 57.48]	35.12 (12.64) [33.73, 36.52]	34.47 (10.52) [33.35, 35.58]
Character Clarity Scale	—	4.72 (0.79) [4.63, 4.80]	4.86 (0.83) [4.77, 4.95]
Characteristic	*n* (%)	*n* (%)	*n* (%)
Education[a]			
0	156 (51.15%)	153 (48.11%)	150 (43.35%)
1	149 (48.85%)	165 (51.89%)	196 (56.65%)
Gender			
Female	190 (62.30%)	165 (52.22%)	217 (63.27%)
Male	115 (37.70%)	151 (47.78%)	126 (36.73%)
Race			
Non-White	28 (9.18%)	57 (17.92%)	66 (19.08%)
White	277 (90.82%)	261 (82.08%)	280 (80.92%)

Note. ART = Author Recognition Test; CI = confidence interval; RMET = Reading the Mind in the Eyes Test; *SD* = standard deviation.
[a]Education: 0 = less than an undergraduate degree; 1 = at least an undergraduate degree.

Table 3. Results of Hypothesis Tests for the RMET.

		df	*F*	*p*	η_p^2 95% CI
Experiment 1	Condition	1,247	0.53	.465	.002 [.000, .027]
	ART	1,247	45.97	<.001	.156 [.081, .236]
	Condition × ART	1,247	5.03	.025	.020 [.000, .066]
Experiment 1 (Not registered)	Condition	1,243	0.65	.421	.002 [.000, .030]
	ART	1,243	35.54	<.001	.127 [.058, .205]
	Condition × ART	1,243	5.38	.021	.021 [.000, .069]
Experiment 2	Condition	1,314	0.01	.924	.000 [.000, .006]
	ART	1,314	31.58	<.001	.091 [.039, .154]
	Condition × ART	1,314	0.01	.918	.000 [.000, .007]
Experiment 3	Condition	1,342	4.91	.027	.014 [.000, .048]
	ART	1,342	19.19	<.001	.053 [.016, .104]
	Condition × ART	1,342	0.69	.408	.002 [.000, .021]

Note. ART = Author Recognition Test; RMET = Reading the Mind in the Eyes Test.

Table 4. Least Squares Means by Condition for the RMET.

		Literary Fiction	Popular Fiction
Experiment 1		25.56 [24.91, 26.21]	25.13 [24.40, 25.86]
	High ART (+1 *SD*)	27.69[a] [26.78, 28.60]	26.20[a] [25.21, 27.19]
	Low ART (−1 *SD*)	23.17 [22.23, 24.10]	23.93 [22.84, 25.02]
Experiment 1 (Not registered)		25.74 [25.09, 26.38]	25.27 [24.55, 25.98]
	High ART (+1 *SD*)	27.65[a] [26.76, 28.55]	26.11[a] [25.14, 27.08]
	Low ART (−1 *SD*)	23.55 [22.62, 24.48]	24.30 [23.23, 25.38]
Experiment 2		26.63 [25.95, 27.30]	26.58 [25.86, 27.30]
Experiment 3		27.03[a] [26.47, 27.60]	26.12[a] [25.55, 26.70]

Note. 95% confidence intervals are presented in brackets. ART = Author Recognition Test; RMET = Reading the Mind in the Eyes Test.
[a]Means in the same row sharing a superscript differ significantly.

Table 5. Cohen's *d* Effect Sizes for Main Effects of Condition.

	RMET	MJT	CCS
Experiment 1	0.092 [−0.155, 0.339]	0.367 [0.114, 0.613]	—
Experiment 1 (Not registered)	.123 [−.147, .352]	.415 [.158, .662]	—
Experiment 2	.011 [−.208, .231]	−.200 [−.481, 0.022]	.265 [.042, .484]
Experiment 3	.239 [.026, .449]	−.213 [−.423, −.00]	.235 [.022, .445]

Note. 95% confidence intervals are presented in brackets. For the "small telescopes" analysis of the effect of condition on the RMET, $d_{33\%} = .203$. CCS = Character Clarity Scale; MJT = moral judgment task; RMET = Reading the Mind in the Eyes Test.

Table 6. Results of Hypothesis Tests for the MJT.

		df	*F*	*p*	η_p^2 95% CI
Experiment 1	Condition	1,247	8.32	.004	.032 [.003, .086]
	ART	1,247	45.83	<.001	.156 [.081, .236]
	Condition × ART	1,247	0.07	.792	.000 [.000, .017]
Experiment 1 (Not registered)	Condition	1,242	10.43	.001	.041 [.006, .098]
	ART	1,242	31.16	<.001	.114 [.048, .189]
	Condition × ART	1,242	0.01	.932	.000 [.000, .008]
	Baseline judgment	1,242	12.02	<.001	.047 [.008, .107]
Experiment 2	Condition	1,313	3.13	.078	.009 [.000, .042]
	ART	1,313	22.32	<.001	.066 [.022, .124]
	Condition × ART	1,313	0.00	.952	.000 [.000, .004]
	Baseline judgment	1,313	27.01	<.001	.079 [.031, .140]
Experiment 3	Condition	1,341	3.90	.049	.011 [.000, .042]
	ART	1,341	24.59	<.001	.067 [.024, .122]
	Condition × ART	1,341	0.23	.631	.000 [.000, .016]
	Baseline judgment	1,341	23.30	<.001	.064 [.022, .118]

Note. ART = Author Recognition Test; MJT = moral judgment task.

Table 7. Least Squares Means by Condition for the MJT.

	Literary Fiction	Popular Fiction
Experiment 1	1.62[a] [1.38, 1.87]	1.08[a] [0.81, 1.36]
Experiment 1 (Not registered)	1.67[a] [1.43, 1.91]	1.07[a] [0.80, 1.34]
Experiment 2	1.56 [1.34, 1.79]	1.86 [1.62, 2.10]
Experiment 3	1.35[a] [1.12, 1.58]	1.68[a] [1.44, 1.91]

*Note.*95% confidence intervals are presented in brackets. MJT = moral judgment task.
[a]Means in the same row sharing a superscript differ significantly, $p < .05$.

Table 8. Results of Hypothesis Tests for the CCS.

		df	*F*	*p*	η_p^2 95% CI
Experiment 2	Condition	1,314	5.53	.019	.017 [.000, .055]
	ART	1,314	0.29	.588	.000 [.000, .019]
	Condition × ART	1,314	1.61	.206	.005 [.000, .031]
Experiment 3	Condition	1,342	4.75	.029	.013 [.000, .047]
	ART	1,342	0.14	.703	.000 [.000, .014]
	Condition × ART	1,342	3.32	.069	.009 [.000, .039]

Note. ART = Author Recognition Test; CCS = Character Clarity Scale.

Table 9. Least Squares Means by Condition for the CCS.

	Literary Fiction	Popular Fiction
Experiment 2	4.62[a] [4.50, 4.74]	4.83[a] [4.70, 4.96]
Experiment 3	4.77[a] [4.65, 4.89]	4.96[a] [4.84, 5.09]

Note. 95% confidence intervals are presented in brackets. CCS = Character Clarity Scale.
[a]Means in a row sharing the same superscript differ significantly.

for the subsequent experiments. As shown in Table 6, results are similar for Experiment 1 using the preregistered and modified analysis plans. However, results varied substantially across studies, with a moderate significant effect in the expected direction found in Experiment 1, a nonsignificant effect in the opposite direction observed in Experiment 2, and a small significant effect in the unexpected direction found in Experiment 3 (see Table 7 for least squares means).

CCS

The CCS was used in Experiment 2 and 3, and it was analyzed in the same way as the RMET: In a GLM with experimental condition, ART scores and their interaction entered as factors. Results from both experiments revealed a significant main effect of condition (see Table 8 for tests of effects and

Table 10. Literary and Genre Author Familiarity as Predictors of RMET and MJT Performance.

			df	*t*	*p*	β 95% CI
RMET	Experiment 1	Literary—ART	1,243	3.89	<.001	.413 [.203, .622]
		Genre—ART	1,243	−0.38	.707	−.039 [−.248, .168]
		Guesses—ART	1,243	−2.24	.026	−.132 [−.249, −.015]
	Experiment 2	Literary—ART	1,314	3.89	<.001	.305 [.150, .460]
		Genre—ART	1,314	0.22	.823	.017 [−.137, .172]
		Guesses—ART	1,314	−1.63	.104	−.087 [−.192, .018]
	Experiment 3	Literary—ART	1,342	3.46	<.001	.264 [.114, .414]
		Genre—ART	1,342	0.15	.884	.011 [−.138, .160]
		Guesses—ART	1,342	−2.15	.032	−.115 [−.222, −.009]
MJT	Experiment 1	Literary—ART	1,242	2.92	.003	.306 [.099, .512]
		Genre—ART	1,242	0.22	.828	.022 [−.184, .230]
		Guesses—ART	1,242	−1.34	.180	−.079 [−.197, .037]
		Baseline judgment	1,242	3.18	.001	.193 [.073, .313]
	Experiment 2	Literary—ART	1,313	3.19	.001	.243 [.093, .393]
		Genre—ART	1,313	0.07	.948	.005 [−.147, .158]
		Guesses—ART	1,313	−0.71	.475	−.037 [−.139, .065]
		Baseline judgment	1,313	5.23	<.001	.280 [.174, .385]
	Experiment 3	Literary—ART	1,341	2.17	.030	.160 [.014, .305]
		Genre—ART	1,341	1.52	.128	.111 [−.032, .256]
		Guesses—ART	1,341	−1.42	.157	−.074 [−.177, .028]
		Baseline judgment	1,341	4.73	<.001	.241 [.140, .341]

Note. ART = Author Recognition Test; RMET = Reading the Mind in the Eyes Test; MJT = moral judgment task.

Table 9 for least squares means), and a marginal interaction with ART was observed in Experiment 3. Scores on the CCS were significantly higher in the popular genre fiction condition than in the literary fiction condition for participants with high (1 *SD* above the mean; $t = 2.82$, $p = .005$, and $d = .304$) and average ART scores ($t = 2.18$, $p = .029$, and $d = .235$), except among participants with low familiarity with fiction (1 *SD* below the mean; $t = 0.25$, $p = .801$, and $d = .027$). Despite the interaction in Experiment 2 not approaching significance, a similar but attenuated pattern emerged, with popular genre fiction characters rated as significantly more clear than literary fiction characters except among participants with low ART scores. A series of unregistered exploratory analyses revealed no correlation between CCS scores and RMET ($ps > .282$) or between CCS scores and MJT performance ($ps > .662$) in either experiment.

Additional analyses

Replication and extension of Kidd and Castano *(2017a).* The present data afford the opportunity to replicate and extend correlational findings concerning the relations of familiarity with literary and genre authors included on the ART and performance on the RMET (Kidd & Castano, 2017a). Literary and genre fiction familiarity scores were calculated as described in Kidd and Castano (2017a) and square-root transformed. Following Kidd and Castano (2017a), guessing scores on the ART were included as a covariate in the models regressing the ToM measures on literary and genre familiarity scores (see Table 10). In the model testing effects on the MJT, baseline judgment scores were also included as a covariate. For both ToM measures, familiarity with literary authors positively predicted performance, but familiarity with genre authors did not.

Discussion

The primary hypothesis, that reading literary fiction would improve RMET performance, received mixed support. Applying Simonsohn's (2015) "small telescopes" method for evaluating the three attempts to replicate Experiment 5 of Kidd and Castano (2013) characterizes two of them as uninformative and one of them as an informative replication. The successful replication here is also consistent with a recently published replication (van Kuijk et al., 2018) that yielded an effect size (Cohen's $d = .358$) similar to that in the original study (Cohen's $d = .33$) and with a 95% CI that includes the effect size observed in the third experiment reported here. Comparisons with similar replication projects (e.g., Panero et al. 2016; Samur et al., 2018) are complicated because they are not direct replications, and some of the deviations from the original methods directly challenge their validity (see Kidd & Castano, 2017b). However, a reanalysis of Panero et al. (2016) yielded a significant difference between the literary and popular genre fiction conditions with an effect size (Cohen's $d = .267$) above $d_{33\%}$ and within the range of the other effect sizes from significant replications (Kidd & Castano, 2017b). The moderately small effects found in the successful replications are consistent with the moderately small effect estimated in the original experiment, indicating that its reliable detection requires large samples and careful controls.

A secondary goal of these experiments was to test whether reading literary fiction would increase the attention paid to

actors' intentions when making moral judgments about their actions. On this task, the MJT results were especially varied across experiments. This may be a consequence of poor validity resulting from using an abbreviated version of the MJT, fatigue (since the MJT was always measured after the RMET), or a true failure of reading literary fiction to influence moral judgments.

Another goal of these experiments was to test the hypothesis that readers would perceive characters in popular fiction as more predictable and stereotypic than those in literary fiction. Consistent with this expectation, both experiments revealed a significant effect of reading condition on CCS scores. The marginal interaction of condition and ART in Experiment 3 further suggested that this effect was most reliable among more experienced readers. Examination of the CCS scores across conditions in both Experiments 2 and 3 suggests that less experienced readers found literary characters a bit more clear than the more experienced readers, and the popular fiction characters slightly less clear. Experienced readers may be more adept at distinguishing between nuanced and simple characters or more quickly adopt different roles as readers based on inferences about the text's genre (Gavaler & Johnson, 2018). However, this finding is not well supported by the present data, and it should be taken primarily as an impetus for further research.

Given the exploratory nature of the CCS, we did not have clear expectations regarding its relations with the RMET or MJT. However, the measure did not reliably correlate with either, suggesting that it cannot account for the effect of condition on ToM. From a theoretical perspective, this is surprising: It is the relatively lower clarity of literary characters that is thought to engage ToM. However, it is difficult to be sure that low ratings of clarity necessarily entail the reader also making a correspondingly greater interpretive effort. Examination of the role of motivation and its sources may help clarify when readers make this effort, which is expected to require more intensive ToM processing.

Conclusions

The primary aim of these three studies was to replicate the effect of reading condition on RMET performance observed in Experiment 5 of Kidd and Castano (2013). The "small telescopes" analyses of the effects observed in the three replication studies suggest that two of the replication studies are uninformative, but that the third was successful. This successful replication is consistent with other close replications (Kidd et al., 2016; van Kuijk et al., 2018; cf. Samur et al., 2018) and a reanalysis of Panero et al. (2016; Kidd & Castano, 2017b). More broadly, these experimental effects align well with the documented positive relation between lifetime experience with literary fiction (but not popular genre fiction) and ToM reported by Kidd and Castano (2017a) and replicated in each of the three present studies.

Attempts to detect a similar effect on the MJT yielded contradictory results, possibly because the substantially abbreviated test used in these studies may not be reliable. Future studies using the full MJT as well as other measures thought to invoke ToM processes remain necessary to assess the generality of the effect observed on the RMET.

In addition to the limitations noted above, the focus on replicating Experiment 5 of Kidd and Castano (2013) entailed relying on only six texts as stimuli. A few conceptual replications have used different texts (e.g., Kidd et al., 2016; Pino & Mazza, 2016), and further research of this sort may help establish the generality of the effects. More critical, though, is further inquiry into how discrete text characteristics affect readers (e.g., Kidd et al., 2016; Koopman, 2016).

Fiction is perhaps the most abundant cultural product, and its effects on readers are surely diverse. There is limited evidence, for example, that reading can increase helping behavior (Johnson, 2012) and empathy (Bal & Veltkamp, 2013), reduce prejudice (Johnson, Jasper, Griffin, & Huffman, 2013), foster openness to ambiguity (Djikic, Oatley, & Moldoveanu, 2013), and alter self-perception (Djikic, Oatley, & Carland, 2012). In some instances, researchers have focused on understanding the diversity of fiction by manipulating aspects of literary style (e.g., Gavaler & Johnson, 2018; Koopman, 2016) or, in correlational studies, working to distinguish among types of genre fiction (e.g., Fong, Mullin, & Mar, 2013).

The present studies can be situated alongside other theoretical (e.g., Culpeper, 2001; Eder et al., 2010) and empirical work (e.g., Kidd & Castano, 2017a; Koopman, 2016) addressing how literary and popular genre fiction engage readers' social cognitive capacities in different ways. In both conditions, participants read works of fiction, with the key difference across conditions being the texts' relative association with popular genres (inclusion in Hoppenstand's [1988] edited anthology) or literary quality (winning the PEN/O. Henry Prize [Furman, 2012]) . Although two experiments yielded inconclusive results, the third shows a significant positive effect of reading literary fiction on the RMET consistent with other experiments using the same methods (Kidd & Castano, 2013; Kidd et al., 2016; van Kuijk et al., 2018). In addition, two experiments yielded strong evidence that characters in literary fiction are perceived as less clear and stereotypic than those in popular fiction. The current evidence is not definitive, but it suggests that researchers interested in how fiction influences social cognition may find it useful to further explore how the methods of characterization in literary and popular fiction evoke different sociocognitive processes.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was in part funded by an Art Works Research Grant of the National Endowment for the Arts (NEA Grant 16-3800-7044).

Supplemental Material

The supplemental material is available in the online version 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: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*, 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*, 241–251. doi:10.1111/1469-7610.00715

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

Bruner, J. S. (1986). *Actual minds, possible worlds*. Cambridge, MA: Harvard University Press. doi:10.2307/357832

Culpeper, J. (2001). *Language and characterisation: People in plays and other texts*. London, England: Longman.

Djikic, M., Oatley, K., & Carland, M. (2012). Genre or artistic merit? The effect of literature on personality. *Scientific Study of Literature*, *2*, 25–36. doi:10.1075/ssol.2.1.02dji

Djikic, M., Oatley, K., & Moldoveanu, M. C. (2013). Opening the closed mind: The effect of exposure to literature on the need for closure. *Creativity Research Journal*, *25*, 149–154. doi:10.1080/10400419.2013.783735

Eder, J., Jannidis, F., & Schneider, R. (2010). Characters in fiction: An introduction. In J. Eder, F. Jannidis, & R. Schneider (Eds.), *Revisionen: Characters in fictional worlds: Understanding imaginary beings in literature, film, and other media* (pp. 3–64). Berlin, Germany: Walter de Gruyter.

Faul, F., Erdfelder, E., Lang, A. -G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fiebich, A., & Coltheart, M. (2015). Various ways to understand other minds: Towards a pluralistic approach to the explanation of social understanding. *Mind & Language*, *30*, 235–258. doi:10.1111/mila.12079

Fong, K., Mullin, J. B., & Mar, R. A. (2013). What you read matters: The role of fiction genre in predicting interpersonal sensitivity. *Psychology of Aesthetics, Creativity, and the Arts*, *7*, 370–376. doi:10.1037/a0034084

Furman, L. (Ed.). (2012). *The PEN/O. Henry Prize Stories 2012*. New York, NY: Anchor.

Gavaler, C., & Johnson, D. (2018). The genre effect: A science fiction (vs. realism) manipulation decreases inference effort, reading comprehension, and perceptions of literary merit. *Scientific Study of Literature*, *7*, 79–108. doi:10.1075/ssol.7.1.04gav

Gelder, K. (2004). *Popular fiction: The logic and practice of a literary field*. London, England: Routledge.

Hakemulder, J. (2000). *The moral laboratory: Experiments examining the effects of reading literature on social perception and moral self-concept*. Philadelphia, PA: John Benjamins. doi:10.1075/upal.34

Hirschfeld, L. (2006). Who needs a theory of mind? In R. Viale, D. Andler, & L. Hirschfeld (Eds.), *Biological and cultural bases of human inference* (pp. 131–160). Mahwah, NJ: Erlbaum.

Hoppenstand, G. (Ed.). (1998). *Popular fiction: An anthology*. New York, NY: Longman.

Johnson, D. R. (2012). Transportation into a story increases empathy, prosocial behavior, and perceptual bias toward fearful expressions. *Personality and Individual Differences*, *52*, 150–155. doi:10.1016/j.paid.2011.10.005

Johnson, D. R., Jasper, D. M., Griffin, S., & Huffman, B. L. (2013). Reading narrative fiction reduces Arab-Muslim prejudice and offers a safe haven from intergroup anxiety. *Social Cognition*, *31*, 578–598. doi:10.1521/soco.2013.31.5.578

Keen, S. (2011). Readers' temperaments and fictional character. *New Literary History*, *42*, 295–314. doi:10.1353/nlh.2011.0013

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

Kidd, D., & Castano, E. (2017a). Different stories: How levels of familiarity with literary and genre fiction relate to mentalizing. *Psychology of Aesthetics, Creativity, and the Arts*, *11*, 474–486. doi:10.1037/aca0000069

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

Kidd, D., Ongis, M., & Castano, E. (2016). On literary fiction and its effects on theory of mind. *Scientific Study of Literature*, *6*, 42–58. doi:10.1075/ssol.6.1.04kid

Koopman, E. M. E. (2016). Effects of "literariness" on emotions and on empathy and reflection after reading. *Psychology of Aesthetics, Creativity, and the Arts*, *10*, 82.

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 Personality*, *40*, 694–712. doi:10.1016/j.jrp.2005.08.002

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

Miesen, H. (2004). Fiction readers' appreciation of text attributes in literary and popular novels: Some empirical findings. *International Journal of Arts Management*, *7*, 45–56.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, *108*, 2688–2692. doi:10.1073/pnas.1011734108

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

Nussbaum, M. C. (1985). "Finely aware and richly responsible": Moral attention and the moral task of literature. *The Journal of Philosophy*, *82*, 516–529. doi:10.2307/2026358

Nussbaum, M. C. (1991). Literary imagination in public life. *New Literary History*, *22*, 877–910. doi:10.2307/469070

Obama, B., & Robinson, M. (2015, November 19). President Obama & Marilynne Robinson: A conversation–II. *New York Review of Books*. Retrieved from http://www.nybooks.com/articles/2015/11/19/president-obama-marilynne-robinson-conversation-2/

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–e54. doi: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:10.3389/fpsyg.2012.00220

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

Rain, M., & Mar, R. (2014). Measuring reading behavior: Examining the predictive validity of print-exposure checklists. *Empirical Studies of the Arts*, *32*, 93–108. doi:10.2190/EM.32.1f

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

Schneider, R. (2001). Toward a cognitive theory of literary character: The dynamics of mental-model construction. *Style*, *35*, 607–639. doi:10.5325/style.35.4.607

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569.

Tamir, D., Bricker, A., Dodell-Feder, D., & Mitchell, J. (2016). Reading fiction and reading minds: The role of simulation in the default network. *Social Cognitive and Affective Neuroscience*, *11*, 215–224. doi:10.1093/scan/nsv114

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*, 7. doi:10.1525/collabra.117

Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D. R., Masala, C., & Preti, A. (2013). The "Reading the Mind in the Eyes" test: Systematic review of psychometric properties and a validation study in Italy. *Cognitive Neuropsychiatry*, *18*, 326–354. doi:10.1080/13546805.2012.721728

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*, 6753–6758. doi:10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*, 8235–8240. doi:10.1073/pnas.0701408104

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *Neuroimage*, *40*, 1912–1920. doi:10.1162/jocn.2009.21137

Zunshine, L. (2015a). From the social to the literary: Approaching Cao Xueqin's *The Story of the Stone* from a cognitive perspective. In L. Zunshine (ed.), *The Oxford handbook of cognitive literary studies* (pp. 176–196). New York, NY: Oxford University Press.

Zunshine, L. (2015b). The secret life of fiction. *Publications of the Modern Language Association of America*, *130*, 724–731. doi:10.1632/pmla.2015.130.3.724

Author Biographies

David Kidd is a postdoctoral fellow at the Harvard Graduate School of Education and studies the impacts of the humanities as associate research director for the Humanities and Liberal Arts Assessment (HULA) project at Project Zero.

Emanuele Castano is a professor at Università degli Studi di Bergamo. His research interests include social identity, intergroup relations, empathy, and the effects of engaging with cultural products on social cognition processes.

Handling Editor: Joseph Simmons

Supplementary Materials and Results

Supplementary Materials

Copies of the pre-registrations, the materials (i.e., PDF copies of the Qualtrics surveys), data, and SAS syntax are available on the OSF database: https://osf.io/yvud8/?view_only=659e06c9805b47eeb89e6dcf2b5fd621

Supplementary Results for Experiment 1

Results including the baseline condition.

RMET. Given the threats to internal validity caused by the failure of random assignment to the baseline condition, the baseline condition was removed from the primary analyses of Experiment 1 in accord with the suggestion of the editor. These results, which should be interpreted tentatively, are presented here (see Table S1 and Table S2). As shown in the tables, the baseline condition RMET scores did not differ from literary or popular fiction condition RMET scores at average ART. At 1 *SD* above the mean, literary condition RMET scores were significantly higher than those in the baseline condition, but the pattern was reversed at 1 *SD* below the mean.

Table S1.

Results of Hypothesis Tests for the RMET

		df	*F*	*p*	η_p^2 95% CI
Experiment 1	Condition	2, 303	0.30	.741	.002 [.000, .017]
	ART	1, 303	22.81	< .001	.070 [.024, .129]
	Condition X ART	2, 303	6.19	.002	.039 [.005, .085]
Experiment 1 (not registered)	Condition	2, 299	0.31	.736	.002 [.000, .017]

	ART	1, 299	18.08	< .001	.057 [.016, .113]
	Condition X ART	2, 299	5.52	.004	.035 [.003, .080]

Table S2.

Least Squares Means by Condition for the RMET

		Literary Fiction	Popular Fiction	Baseline
Experiment 1		25.43	25.06	25.44
		[24.76, 26.09]	[24.31, 25.81]	[24.38, 26.51]
	High ART (+ 1 *SD*)	27.69[ab]	26.20[a]	25.58[b]
		[26.76, 28.62]	[25.19, 27.21]	[23.91, 27.26]
	Low ART (- 1 *SD*)	23.17[a]	23.93	25.31[a]
		[22.21, 24.12]	[22.81, 25.04]	[24.01, 26.60]
Experiment 1 (not registered)		25.60	25.21	25.45
		[24.94, 26.26]	[24.47, 25.95]	[24.39, 26.50]
	High ART (+ 1 *SD*)	27.65[ab]	26.11[a]	25.58[b]
		[26.73, 28.57]	[25.11, 27.11]	[23.93, 27.24]
	Low ART (- 1 *SD*)	23.55[a]	24.30	25.31[a]
		[22.59, 24.51]	[23.20, 25.41]	[24.05, 26.56]

MJT. As shown in Table S3 and Table S4, the MJT scores in the baseline condition did not differ significantly from those in the two reading conditions.

Table S3.

Results of Hypothesis Tests for the MJT

		df	*F*	*p*	η_p^2 95% CI
Experiment 1	Condition	1, 247	8.32	.004	0.367 [0.114, 0.613]
	ART	1, 247	45.83	< .001	.156 [.081, .236]

	Condition X ART	1, 247	0.07	.792	.000 [.000, .017]
Experiment 1 (not registered)	Condition	1, 242	10.43	.001	0.415 [0.158, 0.662]
	ART	1, 242	31.16	< .001	.114 [.048, .189]
	Condition X ART	1, 242	0.01	.932	.000 [.000, .008]
	Baseline Judgment	1, 242	12.02	< .001	.047 [.008, .107]

Table S4.

Least Squares Means by Condition for the MJT

	Literary Fiction	Popular Fiction	Baseline
Experiment 1	1.66[a] [1.40, 1.92]	1.09[a] [0.80, 1.37]	1.35 [0.94, 1.77]
Experiment 1 (not registered)	1.65[a] [1.40, 190]	1.05[a] [0.77, 1.33]	1.41 [1.01, 1.82]

Robust regression.

RMET. As indicated in the pre-registration for Experiment 1, the results of the GLM were confirmed using robust regression, which attenuates the impacts of multivariate outliers. Consistent with the results of the GLM, the main effect of ART was significant ($X^2 = 32.46$, $p < .001$). The main effect of condition was not significant, with scores in the literary fiction condition not differing from those in the popular fiction ($X^2 = 0.62$, $p = .429$) or baseline conditions ($X^2 = 0.07$, $p = .792$), and baseline scores not differing from popular fiction scores ($X^2 = 0.75$, $p = .387$). However, ART scores moderated the contrast between the literary and baseline conditions ($X^2 = 10.33$, $p = .001$) and that between the literary and popular genre fiction conditions ($X^2 = 5.95$, $p = .014$). Follow-up analyses testing these two contrasts at one standard

deviation above the mean revealed that participants in the literary fiction condition obtained marginally higher scores on the RMET than those in the baseline condition ($X^2 = 3.60$, $p = .057$) and significantly higher than those in the popular genre fiction condition ($X^2 = 5.65$, $p = .017$). The baseline and popular genre fiction conditions did not differ ($X^2 = 0.03$, $p = .852$). Estimated at one standard deviation below the mean, the baseline scores were higher than in the literary fiction condition ($X^2 = 7.24$, $p = .007$) but no different than in the popular genre fiction condition ($X^2 = 2.39$, $p = .122$). The literary and popular genre fiction conditions did not differ at one standard deviation below the mean ($X^2 = 0.27$, $p = .602$). Thus, both conventional and robust methods yielded similar pattern of effects, with the exception of the unpredicted higher performance on the RMET in the baseline compared to literary condition at low levels of ART performance.

MJT. The same model described in the manuscript was tested using robust regression. Consistent with the results of the GLM, participants in the literary fiction condition prioritized intent over outcome more than participants in the popular fiction condition ($X^2 = 10.14$, $p = .001$), but not more than those in the baseline condition ($X^2 = 0.75$, $p = .384$). MJT scores in the baseline condition did not differ from those in the popular fiction condition ($X^2 = 2.59$, $p = .107$). There were significant main effects of ART ($X^2 = 7.94$, $p = .004$) and baseline MJT scores ($X^2 = 21.63$, $p < .001$), as in the GLM.