Journal of Personality and Social Psychology

Does Reading a Single Passage of Literary Fiction Really Improve Theory of Mind? An Attempt at Replication

Maria Eugenia Panero, Deena Skolnick Weisberg, Jessica Black, Thalia R. Goldstein, Jennifer L. Barnes, Hiram Brownell, and Ellen Winner

Online First Publication, September 19, 2016. http://dx.doi.org/10.1037/pspa0000064

Journal of Personality and Social Psychology

0022-3514/16/$12.00 http://dx.doi.org/10.1037/pspa0000064

REPLICATION ARTICLE

Does Reading a Single Passage of Literary Fiction Really Improve Theory of Mind? An Attempt at Replication

Maria Eugenia Panero
Boston College

Deena Skolnick Weisberg
University of Pennsylvania

Jessica Black
University of Oklahoma

Thalia R. Goldstein
Pace University

Jennifer L. Barnes
University of Oklahoma

Hiram Brownell and Ellen Winner
Boston College

Fiction simulates the social world and invites us into the minds of characters. This has led various researchers to suggest that reading fiction improves our understanding of others' cognitive and emotional states. Kidd and Castano (2013) received a great deal of attention by providing support for this claim. Their article reported that reading segments of literary fiction (but not popular fiction or nonfiction) immediately and significantly improved performance on the Reading the Mind in the Eyes Test (RMET), an advanced theory-of-mind test. Here we report a replication attempt by 3 independent research groups, with 792 participants randomly assigned to 1 of 4 conditions (literary fiction, popular fiction, nonfiction, and no reading). In contrast to Kidd and Castano (2013), we found no significant advantage in RMET scores for literary fiction compared to any of the other conditions. However, as in Kidd and Castano and previous research, the Author Recognition Test, a measure of lifetime exposure to fiction, consistently predicted RMET scores across conditions. We conclude that the most plausible link between reading fiction and theory of mind is either that individuals with strong theory of mind are drawn to fiction and/or that a lifetime of reading gradually strengthens theory of mind, but other variables, such as verbal ability, may also be at play.

Keywords: replication, theory of mind, mindreading, fiction, reading

Supplemental materials: http://dx.doi.org/10.1037/pspa0000064.supp

Exercising one's mindreading capacities in the context of fictional stories, which tend to focus on interpersonal relationships and psychological states, could lead one to become more empathetic and skilled at mindreading (Keen, 2007; Nussbaum, 2003; Oatley, 2012; Zunshine, 2006). Correlational studies support this argument: Lifetime engagement with fiction, as measured by the Author Recognition Test (ART; Acheson, Wells, & MacDonald, 2008; Stanovich & West, 1989), is positively related to theory of mind, as measured by the Reading the Mind in the Eyes Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), which tests the ability to recognize mental states from photos of a person's eyes (Djikic, Oatley, & Moldoveanu, 2013; Mar, Oatley, Hirsh, dela Paz, & Peterson, 2006; Mar, Oatley, & Peterson, 2009). There is also neural evidence for this connection: Brain areas activated during theory-of-mind tasks are also activated while processing fictional stories (Mar, 2011).

But do people develop their mindreading capacities by reading fiction, or do people with a strong interest in psychological states seek out fictional texts because of fiction's exploration of psychological states? Although the studies cited above are correlational, leaving open this important causal question, Kidd and Castano (2013) reported that a brief, one-time exposure to literary fiction can immediately strengthen social cognition as measured by the

Maria Eugenia Panero, Department of Psychology, Boston College; Deena Skolnick Weisberg, Department of Psychology, University of Pennsylvania; Jessica Black, Department of Psychology, University of Oklahoma; Thalia R. Goldstein, Department of Psychology, Pace University; Jennifer L. Barnes, Department of Psychology, University of Oklahoma; Hiram Brownell and Ellen Winner, Department of Psychology, Boston College.

Data for this study can be found in an Excel file posted on the Open Science Framework (osf.io/83nv2). We thank David Comer Kidd and Emanuele Castano for supplying their data to us and for prompt and gracious responses to many questions. Maria Eugenia Panero, Deena Skolnick Weisberg, and Jessica Black are joint first authors.

Correspondence concerning this article should be addressed to Deena Skolnick Weisberg, Department of Psychology, University of Pennsylvania, 425 South University Avenue, Stephen A. Levin Building, Philadelphia, PA 19104. E-mail: deena.weisberg@psych.upenn.edu

RMET. These researchers randomly assigned participants to read short excerpts of literary texts, popular fiction texts, expository nonfiction texts, or no text. They then assessed participants' theory-of-mind performance and compared it across conditions. Those in the literary fiction condition showed significantly higher RMET scores compared to those in all of the other conditions. This surprising finding attracted much press coverage. *The New York Times* published an article about these findings titled "For Better Social Skills, Scientists Recommend a Little Chekhov" (Belluck, 2013), and a popular blog advised, "Read literary fiction before dates or meetings for social success" (Patkar, 2013).

Based on these results, Kidd and Castano (2013) suggested that there could be an immediate causal connection between reading literary fiction and social–cognitive abilities (see also Oatley, 2016). However, similar studies have found conflicting results. For example, Djikic et al. (2013) failed to find an effect of reading fiction (compared to a nonfiction essay) on the RMET or on the affective empathy scale of the Interpersonal Reactivity Index (Davis, 1983), though they did report that reading fiction positively affected the cognitive empathy scale of the Interpersonal Reactivity Index. Additionally, studies have found that the effects of reading fiction are moderated by individual difference variables such as transportation into the story (Bal & Veltkamp, 2013), affective empathy (Johnson, 2012), and openness to experience (Djikic et al., 2013). Indeed, the only previous attempt to directly replicate Kidd and Castano (2013), which used a within-subjects design, found that participants did score higher after reading literary fiction than nonfiction, but only after statistically controlling for a variety of variables, including narrative transportation (Black & Barnes, 2015). These results suggest that, even in a within-subjects design, the effect may depend on how an individual engages with the specific texts.

Similarly, a recent conceptual replication also found that participants who read literary fiction had higher scores on two tests of mentalizing abilities (a measure of first- and second-order false-belief understanding and a version of the RMET; Pino & Mazza, 2016). However, unlike in Kidd and Castano (2013), these participants read full books and were tested after a 1-week delay, and no effects were found on an additional seven tests of social cognition. These findings paint a more complex picture of the relationship between reading fiction and social–cognitive abilities and suggest that results from intervention studies may be less robust than reported. The small effect sizes in Kidd and Castano (2013; ranging from $\omega_p^2 = 0.01 - 0.05$, corresponding to a 1-point difference on a 36-point scale) support the notion that other variables may be at play and caution against a strong causal interpretation. Further, these prior replication attempts have examined relatively small samples ($N = 60$ to 214), and thus there is a need for replications that pay special attention to the issue of statistical power.

To complicate matters further, Kidd and Castano (2013) report that only literary fiction (not popular fiction) leads to increases in theory of mind. However, the distinction between literary and popular genres of fiction is fuzzy and dynamic (Frow, 2014) and it remains unclear which aspects of literary fiction might be causally responsible. Further, this emphasis on the specific power of literary fiction is puzzling in light of other experiments that obtained effects on social cognition using popular fiction (e.g., Harry Potter; Bal & Veltkamp, 2013; Mutz, 2016; Vezzali, Stathi, Giovannini, Capozza, & Trifiletti, 2015). And other research found higher levels of theory of mind in readers of romance than readers of domestic fiction and science fiction/fantasy (Fong, Mullin, & Mar, 2013).

These suggestive but inconsistent connections between fiction reading and theory of mind led to the current study. In line with recent replication efforts in psychology (Klein et al., 2014; Open Science Collaboration, 2015), we combined data from three independent research groups, each conducting either exact (Group 1) or close conceptual (Groups 2 and 3) replications of Kidd and Castano (2013) to determine whether the original results could be replicated in a between-subjects design. The current study's methods match Kidd and Castano's (2013) more closely than previous attempts and provide more power to detect this effect.

Kidd and Castano (2013) is an important study to replicate because it was often cited in the popular press, as noted above, and thus may begin to be accepted as conventional wisdom. Further, these findings have important implications for potential ways to improve theory of mind, and their robustness should be investigated before implementing any large-scale interventions.

Method

The data reported in this article represent the combined efforts of three research groups, each of which conducted their studies independently and were independently responsible for funding and implementing their designs. Because of this, some measures and conditions varied across the different data sets generated by these groups (see below and Supplemental Materials for details). However, all research groups used the same texts as Kidd and Castano (2013) and included the same key elements of Kidd and Castano's (2013) design, allowing us to combine these separate replication efforts into a single study and thereby boost our power to find effects. Specifically, all participants either read passages of literary fiction, popular fiction, or nonfiction, or they read nothing at all. They were then given the RMET to determine whether literary fiction led to improved theory-of-mind performance compared with other conditions. All participants also completed the ART as a measure of lifetime exposure to fiction.

Research Group 1 completed an exhaustive exact replication of the five experiments described in Kidd and Castano (2013), using Amazon Mechanical Turk (mTurk) workers and Qualtrics software, as they did. However, only Experiments 1, 3, 4, and 5 are reported in this article (see Supplemental Materials for more details on Experiment 2). These are the experiments that included the RMET, the only theory-of-mind variable found by Kidd and Castano (2013) to be positively affected by reading literary fiction. Experiment 1 compared literary fiction to nonfiction, while Experiments 3 and 4 compared literary fiction to popular fiction. Experiment 5 compared literary fiction, popular fiction, and no reading.

Research Group 2 compared literary fiction (texts from Experiment 5), nonfiction (texts from Experiment 1), and no reading, using mTurk workers and Qualtrics software, as in Kidd and Castano (2013). This study was not an exact replication because this group included an instructional manipulation not employed by Kidd and Castano (2013): telling half of the participants that their texts were fiction and the other half that their texts were nonfiction (see Supplemental Materials for more details). However, because there was no effect of this manipulation on participants' scores on

either dependent measure (RMET and ART), these data were combined with the data from the other two groups.

Research Group 3 compared literary fiction and nonfiction, using the same texts as in Kidd and Castano's (2013) Experiment 1 and presenting them using Qualtrics software. Unlike the other two groups and Kidd and Castano (2013), Research Group 3 tested undergraduate participants rather than mTurk workers, making this a conceptual rather than an exact replication.

Participants

The final sample from the three independent research groups consisted of 792 participants (366 males, 422 females, four unreported gender), age range 18 to 76 ($M = 35.39$, $SD = 11.30$). These demographics are comparable to Kidd and Castano's (2013) sample, which was also roughly half female with a mean age of 34.02. Our samples were restricted to the United States; Kidd and Castano (2013) do not report this information for their participants. An additional 510 individuals were recruited but not included in the analyses due to the following exclusion criteria (in order): dropping out before being assigned a text ($n = 46$), reading a text not in the Kidd and Castano (2013) study ($n = 165$), not completing the study ($n = 284$[1]), short or long reading times ($n = 3$), scoring as low outliers on the RMET ($n = 1$), and having a high rate of guessing on the ART ($n = 10$). Kidd and Castano (2013) excluded participants with "inadequately short reading times" for the texts. We also excluded participants with inadequately short reading times (which we defined as 3.5 *SD* below the mean of reading time) as well as those with very long reading times (defined as 3.5 *SD* above the mean of reading time), since long times may have indicated that these participants left their screens without reading. We also excluded participants who scored beyond 3.5 *SD* below the mean on the RMET and the ART, as did Kidd and Castano (2013). Additional details about exclusion procedures and how many participants were excluded from each research group and condition are reported in the Supplemental Materials (Tables S5 and S6).

All subjects were recruited according to procedures approved by the Institutional Review Board at their respective universities. As in Kidd and Castano (2013), mTurk participants were either compensated $2.00 (Research Group 1, Experiments 1, 3, and 4; Research Group 2) or $3.00 (Research Group 1, Experiment 5). Participants from Research Group 3 were recruited from the undergraduate psychology participant pool from their university and so were compensated with course credit.

Materials

We used the same texts as in Kidd and Castano (2013); see Table S1 in the Supplemental Materials for a complete list. The stories in the literary fiction condition were selected by Kidd and Castano because they were winners of prestigious awards for literature (e.g., the 2012 PEN/O. Henry Award for short literary fiction). The stories in the popular fiction condition were selected by Kidd and Castano from an edited anthology of popular fiction and represented a range of genres (e.g., science fiction, romance). The stories in the nonfiction conditions were selected by Kidd and Castano from *Smithsonian Magazine*. These stories report facts about natural and historical topics and do not include biographical narratives about people.

Procedure

Participants were randomly assigned to one of four conditions: literary fiction, popular fiction, nonfiction, or no reading (see Table S1 in the Supplemental Materials); participants in the no-reading control condition did not receive a text. After reading their assigned text (or not reading, for those in the no-reading condition), participants were assessed with two measures: the RMET (Baron-Cohen et al., 2001) and the ART (Acheson et al., 2008; updated from Stanovich & West, 1989). Participants also completed a demographic questionnaire about their age, gender, ethnicity, and highest level of education attained. Additional measures, texts, and conditions were used in different research groups; these are described in the Supplemental Materials.

The RMET, an advanced affective theory-of-mind task, assesses accuracy in mental state and emotion perception. This is a widely used test of the ability to infer a mental state based on an individual's facial expression. This test consists of 36 faces taken from pictures in magazines and edited to reveal only the area between the eyebrows and the bridge of the nose. Each picture is accompanied by four adjectives (e.g., skeptical, joking). Research Group 2 used a version that included a brief definition of each adjective, and Research Group 3 printed out the list of adjectives and definitions from Baron-Cohen et al. (2001) and provided a copy next to each participant's computer. Participants are asked to choose which of these words best describes what the person in the picture is thinking or feeling, and hence is considered both a cognitive and affective theory-of-mind test. Scores are computed by summing the number of correct identifications of expressed emotions.

The ART provides a control for the impact of lifetime exposure to fiction. This test measures familiarity with authors of both popular and literary fiction. It presents a list of 130 names, half of which are authors of works of fiction, and half of which are foils. Participants are told to check only ones that they know for sure are authors, because there is a penalty for guessing. Scores are calculated by subtracting the number of nonauthors selected (the guessing score) from the number of authors identified. Kidd and Castano (2013) used the square root of ART scores in their analyses. Applying this transform reduced a positive skew in the shape of the distribution of ART scores. To apply the same transform to our data, which included negative values as low as -4, we added a constant of 4 to all scores prior to taking the square root.

Analyses and Results

The analyses used by Kidd and Castano (2013) to test the effects of different reading material on RMET scores were problematic in that their fixed effects analysis of covariance (ANCOVA) models ignored potential random effects of stimuli, that is, the different texts used in a single reading condition. In all experiments, there

[1] These excluded participants quit the study without completing one or both of our main measures (RMET and ART), so even those with partial data could not fruitfully be included in our analyses. This rate of quitting is admittedly high; however, these studies asked participants for a considerable time commitment to read the stories and complete the measures, making this study much longer than those more commonly posted on mTurk. Importantly, participants were excluded at roughly equal rates from each condition (see Table S6 in the Supplemental Materials), and our rates of excluding participants for other reasons are comparable to Kidd and Castano's (2013).

were three of each kind of text (literary fiction, popular fiction, nonfiction) and participants were randomly assigned to read one of these texts per condition. Not including random effects due to variation in texts as well as to variation in participant response to texts can lead to upward biasing of the F statistic and alpha inflation (see Judd, Westfall, & Kenny, 2012) and thereby increases the likelihood of false positive findings. Especially with relatively small samples of texts within condition and fixed effects models, there was an increased probability of finding significant effects even if the true effects were very small or null.

We used mixed, or multilevel, models to analyze the data (Judd et al., 2012; see also Maxwell & Delaney, 2004; Singer, 1998). Mixed models account for heterogeneity of variances and covariances, thus permitting the researcher to model effects due to random variables, in this case the effects of stimuli (texts used for each condition), experiment, and research group. Because we only used one experiment per research group in each comparison, we effectively had two random variables: stimuli (text) and experiment (the specific experiment carried out by the corresponding research group). As such, text was nested within experiment which was nested within condition. Incorporation of the random effects in the model avoids the upward bias of significance tests that can occur in less appropriate analysis of variance or ANCOVA approaches (Judd et al., 2012). In each comparison, we entered text as a random variable, nested within experiment within condition (e.g., popular vs. literary fiction). Condition, ART scores, and their interaction were entered as fixed variables in all analyses. To meet the assumptions of normality for mixed models, RMET scores were transformed prior to analyses to correct a negative skew (untransformed means reported to facilitate interpretation). Effect size d for the comparison between means was calculated with adjusted mean difference in the numerator and root mean square error for the model in the denominator. SAS 9.4 PROC MIXED procedure (restricted maximum likelihood estimation; Kenward–Roger degrees of freedom) was used for all mixed models.[2] Table 1 presents sample sizes for each condition for all three comparisons.

Power Analyses

To explore issues related to statistical power, we begin with effect size estimates based on data from Kidd and Castano (2013). Specifically, Table 2 in Kidd and Castano (2013) provides means and standard deviations for each condition in each of their experiments. We used these values, along with sample sizes from their supplementary materials, to calculate Cohen's d for each comparison. Finally, we used G*Power to estimate sample sizes needed to obtain power values for one-tailed tests with $\alpha = .05$ and $1 - \beta = .85$, and also for $1 - \beta = .95$ (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007; see also Simonsohn, 2015).

Table 1
Sources of Data for Each Analysis

Analysis of variance	Condition	Research Group (experiment)	n
Literary vs. nonfiction	Literary fiction	1 (1); 2, 3	191
	Nonfiction	1 (1); 2, 3	109
Literary vs. popular fiction	Literary fiction	1 (3, 4, 5)	151
	Popular fiction	1 (3, 4, 5)	152
Literary vs. no reading	Literary fiction	1 (5); 2	180
	No reading	1 (5); 2	189

Table 2
Reading the Mind in the Eyes Test (RMET) and Author Recognition Test (ART) Scores by Condition and Overall Unadjusted Means for the Current Study and Kidd and Castano (2013), and Zero-Order Pearson's Correlations Between the Two Variables Overall and by Condition

		RMET		ART		
Study	N	M	SD	M	SD	r
Current study						
Grand mean	792	26.28	5.96	17.19	13.31	.47
Literary fiction	342	26.24	5.74	17.96	13.93	.45
Nonfiction	109	27.07	5.12	19.10	13.59	.38
Popular fiction	152	26.05	7.01	17.20	13.08	.45
No reading	189	26.06	5.90	14.67	11.84	.55
Kidd and Castano (2013)						
Grand mean	584	25.18	4.66	21.67	13.92	.26
Literary fiction	225	26.13	4.10	22.81	14.57	.26
Nonfiction	43	23.35	5.18	18.88	13.79	.52
Popular fiction	183	24.50	4.92	21.92	13.52	.22
No reading	133	25.09	4.70	20.30	13.27	.26

Note. RMET and ART scores were transformed to correct for skew prior to correlational analyses. Untransformed means and standard deviations reported. All correlations significant at $p < .01$. Means reported in the text for the primary analyses may differ because the samples were not always the same: In the table, data are pooled across experiments for all those who read the relevant narrative. In the text, only those from experiments that contained both conditions were used.

The literary fiction versus nonfiction comparison (Kidd and Castano's Experiment 1, $N = 86$) has Cohen's $d = .51$, a medium size effect, so a total sample of $N = 114$ is needed for $1 - \beta = .85$, and $N = 170$ is needed for $1 - \beta = .95$. Our comparison has $N = 300$.

For the literary fiction versus popular fiction comparison, Cohen's d was .51, .52, and .30 for Kidd and Castano's Experiments 3 ($N = 69$), 4 ($N = 72$), and 5 ($N = 224$), respectively ($N = 365$ total). Averaging together these three Cohen's d values weighted by the respective sample sizes in the studies yields an average value of .37. The sample size needed for $1 - \beta = .85$ is $N = 202$, and for $1 - \beta = .95$ is $N = 302$. Our comparison has $N = 303$.

For the literary fiction versus no-reading comparison (Experiment 5, $N = 249$), Cohen's $d = .24$. The sample size needed for $1 - \beta = .85$ is $N = 502$, and for $1 - \beta = .95$ is $N = 754$. Our comparison has $N = 369$.

The results we report below for our replication are thus based on sample sizes that exceed what is needed both for good power (i.e., $1 - \beta = .85$) and for excellent power (i.e., $1 - \beta = .95$) in all but the no-reading comparison. Additionally, we note that replication studies have often picked an N to attain a desired high power level based on the effect size estimated in the original study. However, this approach is flawed because publication bias means that the original effect size

[2] We also analyzed our data as Kidd and Castano (2013) did, using ANCOVAs to test for effects of reading condition on RMET scores with ART scores and their interaction with condition as covariates, but with no nesting of text within condition. As in the current analyses, these tests revealed no effects of condition but did find significant relations between RMET and ART scores.

is likely overestimated, resulting in underpowered replication attempts. Simonsohn (2015) discusses this issue and recommends as a rule of thumb that replication studies should have an *N* approximately 2.5 times that of the original study, to have high power to support a potential conclusion that the effect is actually small. Of the three comparisons considered here, our sample size for the first (literary fiction vs. nonfiction) meets this more demanding criterion, while for the other two comparisons we meet the more traditional criterion of having sample sizes comparable to or somewhat larger than the original *N*.

Literary Fiction Versus Nonfiction

All three research groups contributed data for the comparison between literary fiction and nonfiction. Research Group 1 contributed data from their Experiment 1, which assigned participants to read one of the literary fiction texts or one of three nonfiction texts used by Kidd and Castano. In the experiments from Research Groups 1 and 2, participants were randomly assigned to read one of six texts, three literary short stories and three nonfiction articles; the two groups used six different short stories but the same three nonfiction pieces. Research Group 3 had randomly assigned participants to one of two texts for each condition, but only one of each was used in Kidd and Castano (2013), so only the cases that read those texts were used. The fiction text used by Research Group 3 was different from those used by the other two groups, such that the total number of fiction stimuli was seven for this analysis. The nonfiction piece used by Research Group 3 was one of those used by the other groups, such that there were three nonfiction texts for this analysis.

Controlling for ART scores and their interaction with reading condition, and including text as a random variable nested within experiment and condition, RMET scores after reading literary fiction ($M_{adj.} = 26.81$) were no different than RMET scores after reading nonfiction ($M_{adj.} = 27.02$), $F(1, 167) = 0.08$, $p = .775$, $d = 0.07$. ART scores were a significant predictor of performance on the RMET, $F(1, 266) = 44.22$, $p < .001$, $b = 0.19$, but the interaction with condition was not, $F(1, 266) = 0.24$, $p = .625$. None of the random effect intercepts (Text × Experiment × Condition) were statistically significant ($ps > .350$), nor was the variance of the nested random effect ($\sigma^2 = 0.01$, $z = 0.74$, $p = .229$; see Figures 1 and 2).

Literary Fiction Versus Popular Fiction

The comparison between literary and popular fiction included data only from Research Group 1 because this was the only group that included a popular fiction condition. Experiments 3, 4, and 5 from this group were used, in all of which participants were randomly assigned to read literary or popular fiction (in the case of Experiment 5, there was also a no-reading condition; these results are discussed below). In each experiment, participants were ran-

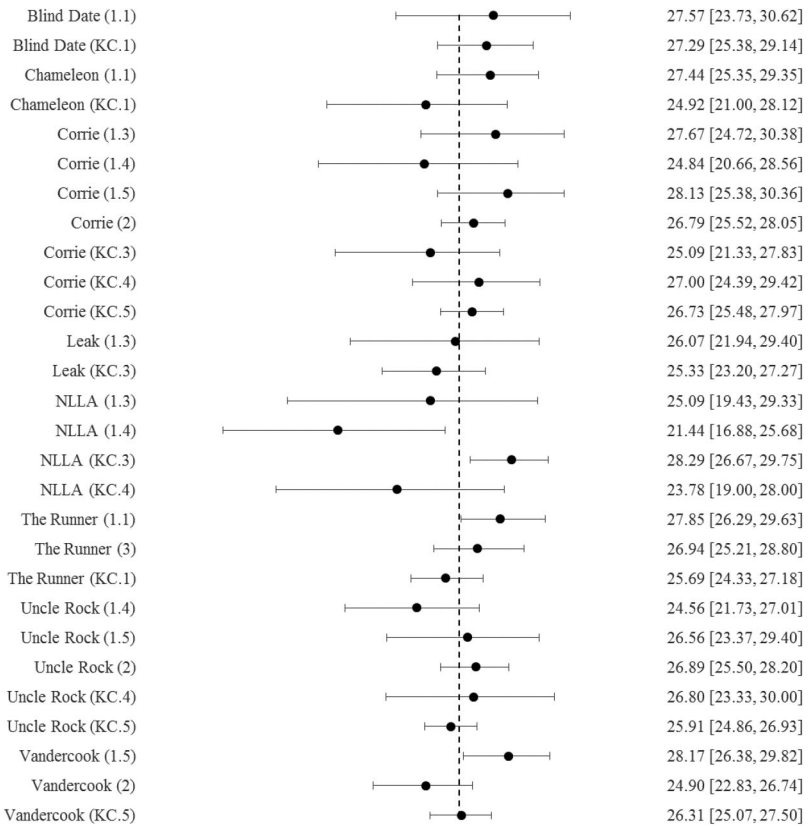

Figure 1. Forest plots of mean Reading the Mind in the Eyes Test scores and 95% confidence intervals for each literary short story. Research group and experiment in parenthesis. Vertical dotted line is at the grand mean ($M = 26.22$). Confidence intervals calculated with bias corrected and accelerated bootstrapping ($N = 5{,}000$). KC = Kidd and Castano (2013); NLLA = Nothing Living Lives Alone.

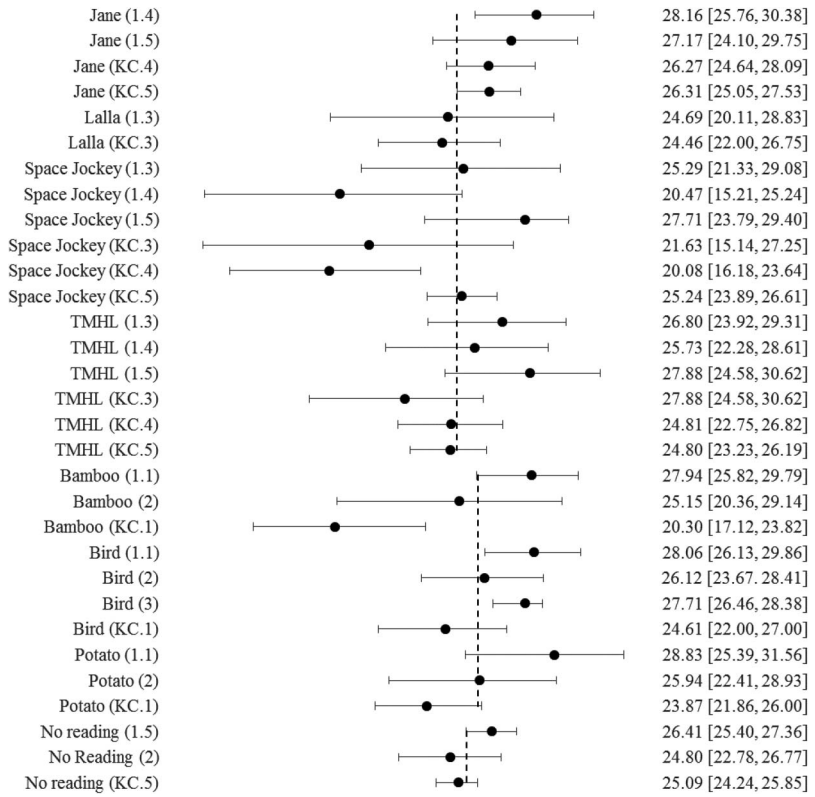

Figure 2. Forest plot for mean Reading the Mind in the Eyes Test scores after each popular fiction story, nonfiction text ("Bamboo Steps Up," "The Story of the Most Common Bird in the World," and "How the Potato Changed the World"), and control condition. Research group and experiment in parenthesis. Vertical dotted lines are at the grand means for popular fiction (M = 25.03), nonfiction (M = 25.85), and control (M = 25.43). Confidence intervals calculated with bias corrected and accelerated bootstrapping (N = 5,000). KC = Kidd and Castano (2013); TMHL = Too Many Have Lived.

domly assigned to one of six texts, three literary and three popular fiction short stories. Across the three experiments, there were a total of five literary stories and four popular stories (see Supplementary Materials for details).

Controlling for ART scores and their interaction with reading condition (as in Kidd & Castano, 2013), and including text as a random variable (nested within experiment and condition), RMET scores after reading literary fiction ($M_{adj.}$ = 25.78) were no different than RMET scores after reading popular fiction ($M_{adj.}$ = 25.99), $F(1, 191) = 0.83$, $p = .363$, $d = 0.05$. As in the previous analysis, ART scores were a significant predictor, $F(1, 288) = 95.16$, $p < .001$, $b = 0.37$, but their interaction with condition was not, $F(1, 288) = 0.67$, $p = .412$. Similarly, the intercept estimates for each random effects level (Text × Experiment × Condition) were not statistically significant (ps > .630), nor was the variance of the nested random effect ($\sigma^2 = 0.01$, $z = 0.54$, $p = .295$; see Figures 1 and 2).

Literary Fiction Versus No Reading

Research Groups 1 (Experiment 5) and 2 contributed data for this final comparison. Note that the data for those who read literary fiction was also used in the prior two comparisons (literary vs. popular fiction for Research Group 1, and literary vs. nonfiction for Research Group 2). Both research groups used the same stimuli for literary fiction (three texts). In Research Group 1 (Experiment 5), participants had been randomly assigned to read literary fiction, popular fiction, or to a no-reading control. In Research Group 2, participants had been assigned to literary fiction, nonfiction, or no-reading control.

Controlling for ART scores and their interaction with reading condition, and including text as a random variable, nested within experiment and condition, there was no difference in RMET scores for participants who had read literary fiction ($M_{adj.}$ = 26.31) than for those who had read nothing ($M_{adj.}$ = 26.25), $F(1, 53.6) = 0.08$, $p = .775$, $d = 0.01$. Once again, ART scores significantly predicted RMET performance, $F(1, 361) = 139.37$, $p < .001$. There was also a significant ART × Condition interaction, $F(1, 361) = 4.06$, $p = .045$: The relationship between ART and RMET scores was slightly stronger in the no-reading condition ($b = 0.37$) than in the literary fiction condition ($b = 0.26$).[3] Again, neither the estimates for random effect intercepts (ps > .24) nor the variance of the nested random effect ($\sigma^2 = 0.02$, $z = 1.02$, $p = .154$) were

[3] Note that these are unstandardized regression coefficients for a model in which both ART and RMET scores had been transformed. The direction and relative values of the relation is correct, but the absolute value is not.

statistically significant (see Figures 1 and 2), though we note that this comparison may have been underpowered to detect effects at the levels found by Kidd and Castano (2013).

Comparison of Our RMET and ART Scores to Kidd and Castano Data

To determine whether the responses in our sample were similar to what Kidd and Castano (2013) found, we compared our mean performance on the RMET and the ART to theirs. For the RMET, our grand mean (26.28) was significantly higher than theirs (25.18), $t(1,374) = 3.71$, $p < .001$, $d = 0.21$. For the ART, Kidd and Castano reported an untransformed mean (21.67) that was larger than ours (17.19), $t(1,374) = 6.06$, $p < .001$, $d = 0.33$. That is, our sample showed higher overall RMET scores and lower overall ART scores, although these differences themselves do not affect our main findings.

Correlations between ART and RMET scores overall and by condition are reported in Table 2. The overall correlation between RMET and ART scores for the current study is significantly stronger than that found in the data from Kidd and Castano, Fisher's $z = 4.46$, $p < .001$. The correlations in the present study are also stronger for literary fiction ($z = 2.53$, $p = .011$), popular fiction ($z = 2.36$, $p = .018$), and no reading ($z = 3.08$, $p = .002$). The correlation for nonfiction is stronger in the Kidd and Castano sample, but not significantly so ($z = 0.95$, $p = .342$).

In contrast to the nonrobust effect of reading condition on RMET scores, ART scores were associated with RMET scores across conditions for both our data and those of Kidd and Castano (2013). Table 3 presents effect sizes for both the magnitude for the differences between means for each reading condition comparison (using only data from experiments that included assignment to each condition) and the strength of the association between RMET and ART scores with the data from each condition combined across experiments for each set of data. Whereas the immediate effect of reading a short text appears unreliable, these effect sizes provide strong evidence for a relation between lifetime exposure to fiction and performance on the RMET.

Discussion

Kidd and Castano (2013) claimed that reading literary fiction (compared to other kinds of texts) improves theory of mind and reported data consistent with this hypothesis. Here, we combined data from three independent research groups to provide a strong test of this claim. We were unable to find significantly higher RMET scores after reading literary fiction compared to any of the other conditions. In short, we found no support for any short-term causal effects of reading literary fiction on theory of mind.

In contrast, ART scores, which measure lifetime exposure to fiction, were consistently significant predictors of RMET scores across all conditions. The correlation between ART and RMET scores could indicate that (a) reading fiction strengthens theory-of-mind skills over time, (b) individuals with stronger theory-of-mind skills are more drawn to fiction (of any kind) than those with weaker theory-of-mind skills, or (c) both.

Overall, then, these findings caution against concluding that there is an immediate effect of reading fiction on theory-of-mind abilities. Strikingly, a prior conceptual replication found an effect of reading condition using a within-subjects design (Black & Barnes, 2015), suggesting that individual difference variables not measured in the current experiment (or in Kidd & Castano, 2013) may play a role in the relationship between reading and theory of mind. We thus believe that we should move from asking whether reading fiction increases theory-of-mind skills to asking under what circumstances reading may do this, and how, and for whom.

Table 3
Effect Sizes for Each Comparison for the Magnitude of the Differences Between Means for Experimental Condition and Overall Association Between Reading the Mind in the Eyes Test and Author Recognition Test Scores (Pooled Across Experiments That Included the Relevant Conditions)

Comparison	*d*	*r*
Data from current study		
Literary fiction vs. nonfiction	−.08 [−.15, .32]	.44 [.37, .51]
Literary fiction vs. popular fiction	.04 [−.19, .26]	.50 [.41, .58]
Literary fiction vs. no reading	.10 [−.10, .30]	.51 [.45, .58]
Data from Kidd and Castano (2013)		
Literary fiction vs. nonfiction	.56 [.13, .99]	.41 [.23, .56]
Literary fiction vs. popular fiction	.36 [.15, .57]	.24 [.14, .34]
Literary fiction vs. no reading	.25 [−.004, .50]	.25 [.15, .35]

Exploring Potential Moderators and Mediators

The effects of reading on theory-of-mind skills are likely to be moderated by individual difference variables, including personality traits (Djikic et al., 2013) and prior exposure to literature (as in the current studies and Black & Barnes, 2015). Similarly, fiction may only facilitate performance on theory-of-mind tasks for readers who enter the task with a particular range of scores on the RMET: Low scorers may benefit more from a reading intervention. Future work should use pretest/posttest designs to investigate these possibilities. Note that although a recent study on this topic (Pino & Mazza, 2016) did use a pretest/posttest design, these authors' use of different tests at the two time points makes it difficult to determine whether their condition differences at posttest were genuinely due to their intervention.

Another variable that may play a role is verbal cognition. The RMET presents sophisticated vocabulary (inquisitive, flirtatious, etc.) that might render the test easier for people with higher verbal skills. Indeed, Peterson and Miller (2012) report that RMET scores from a sample of university students correlated strongly ($r = .49$) with verbal IQ. Nonverbal measures of theory of mind should be considered.

There may also be important differences between different cohorts of readers. Indeed, not only did mean RMET and ART scores differ between our data and Kidd and Castano's, but the relationship between ART and RMET scores was significantly stronger in our data than in that of Kidd and Castano in all conditions except nonfiction. Further, Kidd and Castano (2013) and Research Groups 1 and 2 recruited mTurk workers; Research Group 3 recruited undergraduates from the psychology participant pool. Undergraduates may be taking courses for which they are expected to read literary fiction and may approach anything resembling assigned reading in a different manner than other participants. MTurk workers, who are paid to engage in experiments, may view the reading process as a means to an end.

The way in which a reader approaches a text may also make a difference. Responses are likely to vary if one is genuinely reading, as opposed to skimming, if one is reading for pleasure rather than to obtain information, or if one is reading with the expectation of having to undergo a series of tests afterward. It is also possible that the act of reading itself primes verbal processing in a way that leads to higher RMET scores (Liberman, 2013). These considerations suggest that it is not reading fiction per se that leads to increased theory of mind, but rather the act of reading in a particular way that creates the appropriate mental preparation for performance on mindreading tasks (see Weisberg, Hirsh-Pasek, Golinkoff, & McCandliss, 2014). The inclusion of validity checks such as memory tests will be crucial for future research to determine why relations between reading and social cognition may sometimes occur.

In addition, in both the current study and in Kidd and Castano (2013), participants were given excerpts of texts and were tested immediately following reading. But if there is a general effect of reading on the ability to infer mental states from faces, this may only appear after more prolonged exposure to texts and potentially only after a delay (see Bal & Veltkamp, 2013, on sleeper effects). Future studies should investigate these possibilities directly.

Another possibility for the null results in the current study is that the texts themselves may not have been the best candidates for allowing participants to exercise their mindreading abilities. Although the random-assignment experimental design used here is considered the gold standard for establishing causal claims, the act of choosing for oneself what to read may affect how well any given story will provide practice for one's empathic abilities. For example, the degree to which one is transported into a narrative may predict how well that story affects one's attitudes or abilities. To examine this hypothesis, we analyzed participants' responses to the Narrative Transportation Scale (NTS) from some of the experiments included here (Experiments 1 and 5 from Research Group 1; Research Group 2 and Research Group 3). The NTS includes several questions designed to gauge the degree to which they were transported into the story (e.g., "I could picture myself in the scenes of the events described in the text"; Green & Brock, 2000). We therefore examined the ability of the NTS to explain variance in the RMET. While average NTS scores varied across Research groups, we found no indication that the NTS could account for significant variation in RMET scores. The correlation between NTS and RMET was $r = -.06$ for Research Group 1, $r = +.03$ for Research Group 2, and $r = +.07$ for Research Group 3, all $ps > .4$ These null results, together with those reported by Kidd and Castano (2013), suggest that NTS may not play a role in shaping participants' scores on the RMET, though future work should examine degree of engagement and other individual differences more carefully to fully determine how they might interact with reading condition.

Additionally, the distinction made by Kidd and Castano (2013) between literary and popular fiction is vague, and no independent test was performed to determine whether, as they suggested, literary fiction "uniquely engages the psychological processes needed to gain access to characters' subjective experiences" (p. 378). Further, the selection criteria for the different conditions were biased in that award-winning literary fiction stories were pitted against popular fiction stories that were not selected systematically.

Finally, although Kidd and Castano (2013) aimed to draw conclusions about entire categories of texts, there was considerable variation among texts in the same category as well as across categories. For example, although they are both categorized as literary fiction, "The Runner" is 2,094 words long and has a Flesch-Kinkaid grade level of 3.9, while "The Vandercook" is 5,609 words long and has a Flesch-Kinkaid grade level of 5.6. Another text classified as literary fiction, "Nothing Living Lives Alone," has a Flesch-Kincaid grade level of 10.2. The nonfiction texts ("Bamboo Steps Up," "The Story of the Most Common Bird in the World," and "How the Potato Changed the World") were more homogenous in Flesch-Kincaid grade level but were substantially more difficult (all approximately Flesch-Kinkaid grade level 10) than all other texts except "Nothing Living Lives Alone." These factors reflect intracategory heterogeneity that makes generalization and interpretation problematic. Future studies will benefit from theoretically motivated definitions of those categories that will, in turn, support investigation of factors such as personality variables and level of engagement with different kinds of texts.

Conclusion

The possibility that reading a single brief passage might immediately improve a reader's social skills is exciting and worthy of investigation. However, after a careful study by three independent research groups based on a large number of observations, we are not confident that reading a short text of any kind can reliably improve theory of mind. Any immediate effect of reading on theory-of-mind abilities is likely to be fragile and depend not only on the individual reader and text, but also the relationship between the two. We are thus skeptical about concluding that reading a brief excerpt of literary fiction improves theory of mind in general. We certainly would not recommend any interventions on the basis of the current body of evidence. Nevertheless, given the universality of storytelling, we believe that narrative serves a deep human need and affects our lives in powerful and lasting ways. Rigorous future work should continue to investigate just how narrative exerts its power.

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. http://dx.doi.org/10.3758/BRM.40.1.278

Bal, P. M., & Veltkamp, M. (2013). How does fiction reading influence empathy? An experimental investigation on the role of emotional transportation. *PLoS ONE, 8,* e55341. http://dx.doi.org/10.1371/journal.pone.0055341

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 42,* 241–251. http://dx.doi.org/10.1111/1469-7610.00715

Belluck, P. (2013, October 3). For better social skills, scientists recommend a little Chekhov. *The New York Times*. Retrieved from http://well.blogs.nytimes.com/2013/10/03/i-know-how-youre-feeling-i-read-chekhov/

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

Converse, B. A., Lin, S., Keysar, B., & Epley, N. (2008). In the mood to get over yourself: Mood affects theory-of-mind use. *Emotion, 8,* 725–730. http://dx.doi.org/10.1037/a0013283

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44,* 113–126. http://dx.doi.org/10.1037/0022-3514.44.1.113

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

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,* 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149

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. http://dx.doi.org/10.3758/BF03193146

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. http://dx.doi.org/10.1037/a0034084

Frow, J. (2014). *Genre (The new critical idiom;* 2nd ed.). London, United Kingdom: Routledge.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79,* 701–721. http://dx.doi.org/10.1037/0022-3514.79.5.701

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. http://dx.doi.org/10.1016/j.paid.2011.10.005

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54–69. http://dx.doi.org/10.1037/a0028347

Keen, S. (2007). *Empathy and the novel*. New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195175769.001.0001

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

Kinderman, P., Dunbar, R., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology, 89,* 191–204. http://dx.doi.org/10.1111/j.2044-8295.1998.tb02680.x

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45,* 142–152. http://dx.doi.org/10.1027/1864-9335/a000178

Liberman, M. (2013). That study on literary fiction and empathy proves precisely nothing. *Slate*. Retrieved from http://www.slate.com/blogs/lexicon_valley/2013/10/29/empathy_gap_don_t_believe_that_widely_reported_study_in_science_about_literary.html

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology, 62,* 103–134. http://dx.doi.org/10.1146/annurev-psych-120709-145406

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. http://dx.doi.org/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. http://dx.doi.org/10.1515/COMM.2009.025

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). New York, NY: Taylor & Francis.

Mutz, D. C. (2016). Harry Potter and the deathly Donald. *PS: Political Science and Politics*. Manuscript in preparation.

Nowicki, S. (2010). *Manual for the Receptive Tests of the Diagnostic Analysis of Nonverbal Accuracy 2*. Atlanta, GA: Department of Psychology, Emory University.

Nussbaum, M. (2003). *Upheavals of thought: The intelligence of emotions*. Cambridge, United Kingdom: Cambridge University Press.

Oatley, K. (2012). The cognitive science of fiction. *Wiley Interdisciplinary Reviews: Cognitive Science, 3,* 425–430. http://dx.doi.org/10.1002/wcs.1185

Oatley, K. (2016). Fiction: Simulation of social worlds. *Trends in Cognitive Sciences, 20,* 618–628. http://dx.doi.org/10.1016/j.tics.2016.06.002

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* 910–911.

Patkar, M. (2013, October 12). Read literary fiction before dates or meetings for social success. *LifeHacker*. Retrived from http://lifehacker.com/read-literary-fiction-before-dates-or-meetings-for-soci-1444007471

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. http://dx.doi.org/10.3389/fpsyg.2012.00220

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

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

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559–569. http://dx.doi.org/10.1177/0956797614567341

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23,* 323–355. http://dx.doi.org/10.2307/1165280

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

Vezzali, L., Stathi, S., Giovannini, D., Capozza, D., & Trifiletti, E. (2015). The greatest magic of Harry Potter: Reducing prejudice. *Journal of Applied Social Psychology, 45,* 105–121. http://dx.doi.org/10.1111/jasp.12279

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

Weisberg, D. S., Hirsh-Pasek, K., Golinkoff, R. M., & McCandliss, B. D. (2014). Mise en place: Setting the stage for thought and action. *Trends in Cognitive Sciences, 18,* 276–278. http://dx.doi.org/10.1016/j.tics.2014.02.012

Zunshine, L. (2006). *Why we read fiction: Theory of mind and the novel*. Columbus, OH: Ohio State University Press.

Received May 25, 2016
Revision received August 22, 2016
Accepted August 22, 2016 ■