Journal of Experimental Psychology: General

Does a Brief Exposure to Literary Fiction Improve Social Ability? Assessing the Evidential Value of Published Studies With a p-Curve

Joshua A. Quinlan, Jessica K. Padgett, Amin Khajehnassiri, and Raymond A. Mar

Online First Publication, October 20, 2022. http://dx.doi.org/10.1037/xge0001302

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001302

Does a Brief Exposure to Literary Fiction Improve Social Ability? Assessing the Evidential Value of Published Studies With a *p*-Curve

Joshua A. Quinlan, Jessica K. Padgett, Amin Khajehnassiri, and Raymond A. Mar
Department of Psychology, York University

Humans have long suspected that stories can help us better understand others, and, indeed, lifelong exposure to narrative fiction does predict better social cognition. Several experiments have attempted to investigate the causal direction of this relationship to see if random assignment to a brief narrative directly improves social cognition. Although these experiments have yielded mixed results, a recent meta-analysis did find a small causal effect of narrative fiction on social cognition. What remains unanswered is whether the published findings reflect questionable research practices or trustworthy evidence. In order to rule out the possibility that this body of work has been meaningfully impacted by selective reporting, we conducted a *p*-curve on the experimental literature on narrative fiction and social cognition. The results of the *p*-curve indicated that this work does indeed have evidential value but that this conclusion is not very robust. Thus, further experimental work on the causal effect of narrative fiction on social cognitive skills is required before substantial conclusions can be drawn.

Keywords: literary fiction, narrative, social cognition, *p*-curve, theory of mind

Humans have long suspected that stories might help us to better understand our peers. In *The Analects*, compiled some 2,000 years ago, Confucius encouraged his followers to "study the Book of Poetry" as it can "be used for purposes of self-contemplation" and "teach the art of sociability" (Confucius, 500 BCE/1861, 17:9). This ancient idea that narrative fiction can help us to better understand ourselves and others has recently received renewed attention from scientists. A wealth of correlational studies have found that long-term exposure to narrative fiction is positively associated with social cognitive skills (e.g., Mar et al., 2006; see meta-analysis by Mumper & Gerrig, 2017). However, the correlational nature of these studies means that we cannot confidently infer that exposure to narrative fiction causes an enhancement of social cognition. In order to clarify the causal nature of this relationship, there have been several attempts to use experiments to uncover whether exposure to narrative fiction improves social cognition (e.g., Kidd & Castano, 2013; Samur et al., 2018). Unfortunately, these experiments have yielded mixed results, with some high-powered replications failing to find any effect (e.g., Panero et al., 2016; cf. van Kuijk et al., 2018). In light of these mixed findings, how can we best identify whether this body of literature provides good evidence for a true effect?

One promising option is meta-analysis. A recent meta-analysis of these experiments concluded that brief exposure to narrative fiction does have a small positive effect on social cognition (Dodell-Feder & Tamir, 2018). Although the results of this meta-analysis indicate an overall effect for these experiments, it cannot tell us whether the findings it summarizes are primarily the result of questionable research practices. These questionable research practices often entail making self-serving decisions during the research process that increase the chances of confirming one's hypotheses: making a Type I error (Simonsohn et al., 2014). Although meta-analyses cannot detect questionable research practices (such as selective reporting), a recently developed statistical method can achieve this goal: the *p*-curve (Simonsohn et al., 2014). Importantly, a *p*-curve of the experiments from one high-profile article on this topic did find that this group of experiments lacked evidential value (van Kuijk et al., 2018). However, a *p*-curve has yet to be conducted on all of the extant experiments on this topic. In the present research, we build upon this previous meta-analytic work, applying a *p*-curve analysis to all of the experimental literature on reading narrative fiction and social cognition in order to rule out the possibility that this body of work has been meaningfully impacted by questionable research practices. Such a demonstration will help increase our confidence in this literature and also help determine whether exposure to narrative fiction improves social cognition.

Narrative Fiction and Social Cognition

Psychologists have defined stories as representations of a series of goal-centered events that are temporally organized and causally related (Rumelhart, 1975). These events typically follow a common structure, comprising an inciting incident, rising action, a resolution,

Joshua A. Quinlan https://orcid.org/0000-0002-6320-0806

This work was supported by the Social Sciences and Humanities Research Council of Canada under Grant 435-2017-1030. Prior to this publication, Raymond A. Mar presented this work to the Institute for Psychology, Cognitive Science, and Gender Research at the University of Freiburg (Germany). The data and additional information for this analysis can be found at https://osf.io/mkynj/ (Quinlan et al., 2021).

Correspondence concerning this article should be addressed to Joshua A. Quinlan, Department of Psychology, York University, Behavioral Sciences Building, 315 4700 Keele Street West, Toronto, ON M3J 1P3, Canada. Email: joshuaq@yorku.ca

and a denouement (Stein & Glenn, 1975). In addition to a common structure, stories also share a focus on the intricacies and vicissitudes of social relations, thereby allowing audiences to mentally simulate complex social interactions, including the mental states of characters and their emotions (Boyd, 2009; Hogan, 2003; Mar & Oatley, 2008). Because understanding narrative fiction requires us to contemplate the cognitive and emotional states of characters, it has been suggested that stories engage the same social cognitive processes we use when interacting with others in the real world (Gerrig, 1993; Zunshine, 2006). By immersing audiences in a simulation of social experiences that encourages them to consider the perspectives of others, it is possible that narrative fiction could help improve social cognition (Mar, 2018a, 2018b), in line with the suggestion by Confucius in *The Analects*.

In the millennia following Confucius's prescription, psychologists have empirically assessed his claim that narrative fiction can help us to better understand one another. Much of this research has focused on two aspects of social cognition: (a) the ability to infer the cognitive and emotional states of others, known as theory of mind or mentalizing (Carruthers & Smith, 1996), and (b) empathy, or feeling the emotions of others (Preston & de Waal, 2002). The relationship between narrative fiction and social cognition emerges in early childhood, with exposure to children's storybooks predicting theory of mind development in 4–5-year-old children (Adrian et al., 2005; Mar et al., 2010). This relationship persists into adulthood, with lifetime exposure to narrative fiction predicting better mentalizing ability in adults (e.g., Fong et al., 2015; Mar et al., 2006). This latter finding appears to be robust, with a recent meta-analysis of 22 correlational studies corroborating a positive association between reading narrative fiction and both mentalizing and empathy (Mumper & Gerrig, 2017).

Experimental Evidence

Although there is substantial correlational evidence that exposure to fiction is positively related to social cognition, these correlational findings cannot be used to support causal inferences regarding their relationship. It may be that people with better social cognitive abilities are more attracted to narrative fiction, or some third variable may cause both an attraction to fiction and improved social cognition. This has led many researchers to attempt experimental demonstrations that reading narrative fiction causes improvements in social cognition as only experiments can support causal inferences. In these experiments, participants typically read either a short piece of fiction (e.g., a short story or excerpt from a novel) or nonfiction (e.g., a newspaper article) and then complete measures of social cognition. One commonly used measure is the Reading the Mind in the Eyes task, which asks participants to identify the emotional state of a person based on a photograph of their eye region (Baron-Cohen et al., 2001). Other studies have used self-report measures of trait empathy (e.g., the Interpersonal Reactivity Index; Davis, 1980), false-belief measures of theory of mind (e.g., Converse et al., 2008), or emotion attribution tasks for fictional characters (e.g., Blair & Cipolotti, 2000). However, the results of these experiments have failed to yield consistent results. In a set of high-profile studies, Kidd and Castano (2013) showed that brief exposure to literary fiction could improve performance on theory of mind tasks,[1] unlike exposure to popular fiction (Studies 2–5) or nonfiction (Study 1). Unfortunately, attempts to replicate this finding have been mixed, with several high-powered direct replications failing to find any effect (Panero et al., 2016; Samur et al., 2018; cf. Kidd & Castano, 2017; Panero et al., 2017) and one successful replication (van Kuijk et al., 2018). In order to resolve these inconsistent results, Dodell-Feder and Tamir (2018) conducted a meta-analysis of experiments (both published and unpublished) that tested whether a brief exposure to narrative fiction could cause an improvement in social cognition (operationalized as theory of mind, empathy, and prosocial behaviors). This meta-analysis comprised 53 effect sizes from 14 articles and showed that reading fiction led to a small, but robust, increase in social cognition (g = .15, 95% confidence interval [CI; .02, .29]). The authors found no evidence that this effect was moderated by any study characteristics, such as the type of sample or outcome measure. Finally, using funnel plots, the authors also found no evidence of publication bias, although the average effect size was smaller among the unpublished studies (g = .08) than the published ones (g = .19).

Questionable Research Practices

The results of this meta-analysis by Dodell-Feder and Tamir (2018) answer many of our questions about the mixed findings for experiments on this topic. It tells us that the literature overall presents evidence for an effect, one that cannot be attributed to moderators or to publication bias. However, it cannot tell us one important thing: whether the results reported are the result of questionable research practices. In other words, do the experiments being summarized possess evidential value? One form of questionable research practices that might be at play is *p*-hacking, in which a researcher capitalizes on the flexibility of various aspects of data analysis and design to produce *p*-values below the alpha threshold for null-hypothesis statistical testing (traditionally set at .05; Simonsohn et al., 2014). Because researchers must make myriad decisions about how to proceed with data analysis, there are many "researcher degrees of freedom" that can be exploited, some of which influence whether or not an effect is observed (Simmons et al., 2011). Example decisions include whether to collect more data, which covariates to include or exclude, or how and when variables should be combined. If these decisions are made based on whether the outcome supports the hypothesis or not (i.e., produces a $p < .05$), then the chance of reporting false positives increases markedly.

For example, consider a researcher who estimates a model and finds a statistically nonsignificant result. If this outcome motivates her to then estimate a new model with an additional covariate, one that does produce a statistically significant result, this new model is likely to be adopted and reported. Estimating new models until one produces a $p < .05$ greatly inflates the odds of finding falsely positive results. Only reporting the statistically significant model then misrepresents the odds that the finding is a false positive.

[1] Kidd and Castano (2013) defined literary fiction as literature that engages the creativity of the reader and requires their imaginative input in order to be properly understood and appreciated. They argued that literary fiction uniquely engages the psychological processes involved in theory of mind and so should enhance the skill, whereas popular fiction only entertains readers and so does not affect the ability. For the present study, however, we are evaluating the claim that reading narrative fiction in general can improve social cognition, without distinguishing between literary and popular fiction.

This approach, among other seemingly innocuous practices, could contribute to a body of work that lacks evidential value. Notably, *p*-hacking is a separate phenomenon from publication bias (addressed by the previous meta-analysis; Dodell-Feder & Tamir, 2018). Publication bias attempts to estimate if there is a substantial "file drawer" of unpublished results that might overturn an estimate of an effect. It does not evaluate whether the results considered might have arisen from *p*-hacking or some other questionable research practice.

The *p*-Curve

The *p*-curve assesses evidential value by plotting the magnitude of *p*-values from a set of related studies (Simonsohn et al., 2014). If the true effect under investigation is null, the distribution of *p*-values should be totally flat, with the same number of *p*-values at each magnitude (e.g., the same frequency of values observed at .01 as at .05). For a true effect, however, the distribution of *p*-values should be right skewed, with more *p*-values occurring at lower values than higher (e.g., more at .01 than at .05). Such a distribution of *p*-values suggests that the collection of studies have good evidential value. In contrast, if a distribution of *p*-values is not right skewed, this suggests the body of work lacks evidential value. Depending on the power of the included studies and the nature of the effect under investigation (i.e., whether it is true or null), *p*-hacking could result in a left-skewed curve or it could flatten the curve.[2] However, the *p*-curve does not test for flatness or left skewness of a curve; it only assesses whether the curve is right skewed. That is, the *p*-curve only tests for whether the body of literature indicates good evidential value; it does not test for evidence of *p*-hacking directly. That said, a distribution of *p*-values that is not right skewed suggests that the true effect may be null or that these findings are influenced by *p*-hacking. Because the traditional cutoff for statistical significance is .05, there is little additional incentive to obtain a *p*-value that is lower than .049, and so *p*-curves affected by *p*-hacking should show a concentration of *p*-values close to .05.[3]

It is worth noting that a *p*-curve is not intended to determine whether an effect under investigation is real. Although a right-skewed curve would be indicative of the existence of a true effect, a curve that is not right skewed is agnostic in this regard (Simmons & Simonsohn, 2017). A curve that is rather flat could indicate that the effect under investigation is null or that a substantial number of low-powered studies were *p*-hacked in the presence of a true effect. A left-skewed curve suggests the presence of substantial *p*-hacking for a null effect, but a curve that is neither left nor right skewed is somewhat ambiguous. Even a left-skewed curve is not necessarily evidence of a null effect, just that the current body of studies lacks evidential value.

The *p*-curve analysis relies on testing a distribution of published, statistically significant *p*-values for right skewness. If the distribution is not observed to have the desired right skew, a test is performed to examine whether the *p*-curve has sufficient power to make a conclusion. This is achieved by testing the similarity between the curve observed and what would be expected if the collection of studies had an average power of 33%. If the distribution is flatter than the curve one would expect with 33% power, one can conclude that the set of studies does not have good evidential value. *P*-curves that are not right skewed but are also not flatter than the expected curve when the average power is 33% are inconclusive and require additional *p*-values to determine whether the studies have evidential value (Simonsohn et al., 2014). This test ensures that underpowered *p*-curves do not mistakenly conclude that a set of studies lacks evidential value (Simmons & Simonsohn, 2017).

Given that the *p*-curve can potentially indicate that a set of studies lacks evidential value, some may be concerned that it could be used to spuriously discredit a true effect. Although this is technically possible, it is very unlikely. The *p*-curve is, by design, highly powered to detect evidential value and very conservative in concluding that a set of studies lacks evidential value (Simonsohn et al., 2014). Using simulated data, the original authors demonstrated that a *p*-curve with as few as 10 *p*-values from properly powered studies (i.e., 80% power) falsely concludes that they lack evidential value less than 0.1% of the time (p. 544). As the number of *p*-values included in the curve increases, this likelihood further decreases. Thus, the *p*-curve is highly unlikely to erroneously conclude that a set of studies lacks evidential value.

In the present study, we used a *p*-curve to determine whether the experimental evidence for a causal effect of reading narrative fiction on social cognition possesses good evidential value or is likely to be the result of questionable research practices. Notably, a *p*-curve analysis conducted on four studies from Kidd and Castano's (2013) seminal paper indicated that this set of studies has low evidential value (van Kuijk et al., 2018). This result, along with the inconsistent findings in the literature, demonstrates the need for a *p*-curve on all of the experimental studies on this topic. In order to avoid any potential biases on our part, we closely followed the methodology laid out in Dodell-Feder and Tamir's (2018) meta-analysis. Specifically, we followed their article selection process, operationalizing social cognition as "any measure testing the processes that underlie how one perceives, interprets, and responds to social information" (Dodell-Feder & Tamir, 2018, p. 1714). As it is highly unlikely that these authors intentionally conducted their meta-analysis in such a way that a future *p*-curve would find some particular outcome, following their process allowed us to conduct this analysis in an unbiased fashion. By determining whether this body of experiments has good evidential value, we hoped to extend our understanding of the effects of narrative fiction and inform directions for future research in this area.

Method

The current study was preregistered prior to data collection, as is recommended for analysis of preexisting data sets (https://aspredicted.org/9es9g.pdf; Mertens & Krypotos, 2019; Weston et al., 2019). As discussed in the preregistration, the *p*-curve analysis consists of several steps: (a) conduct a systematic search for the literature to be analyzed, (b) select the articles to be included in the *p*-curve, (c) collect the relevant statistical results from each article, and (d) input the results into the online calculator and run the *p*-curve. In this section, we provide a detailed

[2] Discussion of this discrepancy can be found in Section 3 of the online supplemental materials for the original *p*-curve article (Simonsohn et al., 2014).

[3] Although there is little incentive to *p*-hack *p*-values far below .05, the *p*-curve is also robust to more ambitious *p*-hacking (Simonsohn, Simmons, & Nelson, 2015).

account of each step. Relevant materials discussed in the methods can be found on the Open Science Framework (OSF; https://osf.io/mkynj/).

Systematic Review

The articles selected for the *p*-curve were collected through a process that borrowed heavily from the Cochrane guidelines for systematic review (Higgins, 2019) and, to a lesser extent, the PRISMA guidelines (Liberati et al., 2009). The search terms and databases used for this process were based on those used in the Dodell-Feder and Tamir (2018) meta-analysis. However, rather than simply pulling the articles that they relied upon, we repeated this search ourselves for the sake of thoroughness and to ensure that any articles published since their analysis would be included. Unlike a typical systematic review for a meta-analysis, but consistent with the guidelines of the *p*-curve, we were only interested in published work. Generally, studies that "work" (i.e., obtain a *p*-value below .05) are published, whereas studies that do not "work" are put in the file drawer (Rosenthal, 1979). Although it is possible that unpublished work has been subjected to questionable research practices, such work has no influence on our understanding of a given phenomenon as it remains largely unknown and is therefore irrelevant to this project.

Following the Cochrane guidelines for a systematic review, we collected all potential articles through a broad database search. We searched PubMed, PsycINFO, and Web of Science for English-language, peer-reviewed articles from academic journals that experimentally estimated the effect of reading narrative fiction on social cognition. Our search terms were taken directly from the aforementioned meta-analysis and included the word "fiction" paired with 10 different search terms related to social cognition: *fiction* AND (*social cognition* OR *social ability* OR *social skill* OR *social perception* OR *theory of mind/theory-of-mind* OR *mentalizing* OR *mind reading* OR *perspective taking* OR *empath** OR *emotion*; Dodell-Feder & Tamir, 2018).

After the search was conducted for each database, metadata that would be relevant for the next steps were gathered from the search results. This included author names, article title, journal name, and abstract. Results from each database were compiled into a single document where duplicates were removed. The final set of articles after this process was 529.

Article Selection

After articles from all three databases were collected and combined and all duplicates were removed, we identified the articles to be included in our analyses using our preregistered inclusion criteria (Higgins et al., 2019). In order to be included, the articles had to (a) be published in a peer-reviewed journal, (b) be written in English, (c) be a true experimental design (between subjects or within subjects) with random assignment to condition, (d) be a comparison between exposure to narrative fiction versus a control (e.g., popular fiction, nonfiction, no reading at all), and (e) include an outcome variable that is a form of social cognition as defined by Dodell-Feder and Tamir (2018; i.e., theory of mind/mentalizing, empathy, and/or prosocial behavior).

A codebook was developed for research assistants who were tasked with reading a subset of the abstracts and noting whether the articles met the inclusion criteria. If the abstract indicated that the article met all the inclusion criteria, the coder would note that the article should be included in the analysis (i.e., it would move forward to Step 3). The abstracts were split into three subsets, each of which was coded by two independent research assistants. In cases where a research assistant was unsure about an article's candidacy for inclusion, this was noted in the coding and the authors made a judgment based on consensus. Similarly, if research assistants disagreed on the inclusion of a specific article, the authors made a final decision based on consensus. Thus, all articles identified for potential inclusion based on their abstracts were considered by two research assistants and all authors, resulting in a final list of 24 articles. After a closer reading of the articles themselves by the authors (rather than just the abstracts), a further eight were removed for failing to meet the inclusion criteria, resulting in a final set of 16 articles to be included in the final analysis. A document containing the coded abstracts can also be found on the OSF page (https://osf.io/mkynj/).

By closely following the operationalization and process employed by Dodell-Feder and Tamir's (2018) published meta-analysis, we circumvented the possibility that our selection of studies is biased in favor of a particular outcome. As suggested by Simmons and Simonsohn (2017) in their *p*-curve of the power posing literature, because our studies will have been chosen "*for* us, not *by* us" (p. 690), it is very unlikely that study selection is guided by how large or small the critical *p*-values are. Additionally, by preregistering our literature review techniques and our inclusion criteria, and by documenting and reporting each selection decision, we ensured that our *p*-curve was both valid and unbiased.

Selecting Relevant Statistical Results

After the final set of studies were selected, we moved on to the third step and selected the statistical results to be included in the *p*-curve analysis, following the guidelines set out by Simonsohn and colleagues (2014). These guidelines mandate the completion and publication of a disclosure table to highlight all the decisions made during this process. This table includes (a) a selection of quoted text from the original article indicating the primary hypothesis relevant to the researchers, (b) the study design, (c) the key statistical result that should be taken, (d) quoted text from the original article that includes the statistical result, (e) the result as calculated by the *p*-curve app, and (f) whether there are any statistics that warrant inclusion in a robustness *p*-curve.

A robustness *p*-curve is an essential component of the *p*-curve, repeating the analysis using alternative test statistics. This ensures that the result of the initial *p*-curve does not change if slightly different, but equally defensible, choices are made when extracting the test statistics. These alternative test statistics arise for two reasons: (a) there is ambiguity in the text as to which test statistic is appropriate for the analysis, or (b) a particular analysis produces two relevant *p*-values from the same sample (NB. for nearly all study designs, only one value from each sample can be included in a single *p*-curve). To produce the robustness *p*-curve, alternative test statistics are collected while selecting test statistics for the main analysis and an additional *p*-curve is run, swapping in these alternative *p*-values.

All articles were examined independently by two authors, who each created their own disclosure table. In cases where different

decisions were made, or if there was uncertainty regarding which hypothesis was of primary interest or which statistical result should be included, the authors discussed the issue and came to a consensus on the most appropriate decision. If total consensus could not be reached, or in ambiguous cases, alternative choices for the test statistic that are not included in the main *p*-curve analysis were noted for inclusion in the robustness *p*-curve described above.

As mentioned, one of the challenges involved in conducting a *p*-curve analysis is determining which statistical test should be selected from a study. This can be especially difficult when there are complicated models that have multiple results all related to the same hypothesis. However, the *p*-curve manual gives clear instructions and justifications for which statistical test to select for each analysis (Simonsohn et al., 2014). For example, in a study that has multiple group comparisons, the omnibus analysis is unlikely to be associated with the main hypothesis. Instead, comparisons between groups are likely the statistical tests of interest, but it may sometimes be unclear *which* group comparison should be included in the *p*-curve. Having anticipated this difficulty, the *p*-curve manual provides guidelines for determining which test statistics to take from a variety of common study designs. Consider a three-condition study design in which one group reads literary fiction (a "treatment" condition), one group reads nonfiction (first control condition), and one group reads nothing at all (second control condition). The *p*-curve authors suggest that the comparison between literary fiction and nonfiction should be included in the main *p*-curve analysis and the comparison between literary fiction and no reading should be included in a robustness *p*-curve. We followed these guidelines closely when selecting test statistics to determine which tests should be included in the main analysis and which should be included in the robustness *p*-curve in order to ensure that our *p*-curve analysis was correctly specified. In addition to our final disclosure table, a list of the statistical tests included in both the main *p*-curve and robustness *p*-curve is included on the OSF.

p-Curve Analysis

Once all relevant test statistics were collected, we used the *p*-curve app to generate a *p*-curve and its associated analyses (www.p-curve.com/app4). The first analysis generated by the *p*-curve app is a test of right skewness. If either the left half of the *p*-curve is right skewed (at $\alpha = .05$) or both the full curve and the left half of the curve are right skewed (at $\alpha = .10$), then the *p*-curve is deemed to have good evidential value (Simonsohn et al., 2015).

P-curves that are not right skewed are the result of either a precise estimate of a very small or nonexistent effect or an imprecise estimate of an effect of any size (Simonsohn et al., 2014). In order to distinguish between these two cases, the *p*-curve is then tested for whether it is flatter than the curve expected from a set of studies that have 33% power. A *p*-curve that is flatter than this 33% power curve indicates that the studies it comprises are either estimating an effect that is nonexistent or are too underpowered to estimate said effect. If either the left half of the *p*-curve is flatter than the 33% power curve (at $\alpha = .05$) or both the full curve and the left half of the curve are flatter than the 33% power curve (at $\alpha = .10$), then the studies included in the *p*-curve are deemed to have inadequate evidential value (Simonsohn et al., 2014, 2015). This result would indicate that the studies included in our *p*-curve do not have good evidential value, are not an accurate estimate of the effect of narrative fiction on social cognition, and are unlikely to replicate. If the result of this flatness test is also null, the *p*-curve itself is underpowered and its results are inconclusive. In this case, additional *p*-values would be needed in order determine the evidential value of these studies (Simonsohn et al., 2014).

Additionally, the *p*-curve app includes a leave-one-out analysis, where the highest and lowest *p*-values are dropped from the analysis and the significance of each test (right skew of the full *p*-curve, right skew of half the *p*-curve, and the 33% power analysis) is recalculated. For example, in the "drop K lowest original *p*-values," the smallest *p*-values are dropped one at a time, and the results of the tests after having removed the most extreme *p*-values are plotted. The purpose of this is to determine the extent to which the results of the *p*-curve are dependent on a few studies or a few extreme *p*-values. Sets of studies whose *p*-curve results remain consistent when extreme results are excluded are more trustworthy than those that change when a few studies are omitted.

Results

Main *p*-Curve

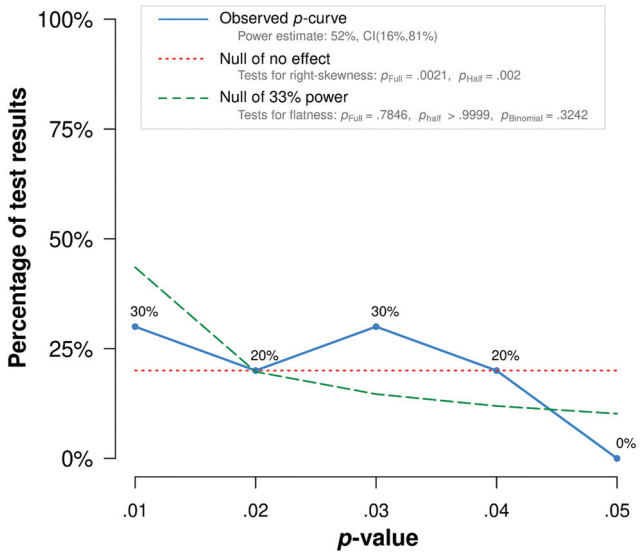
Our disclosure table includes 16 articles, but three could not be used as they failed to report the necessary details about their analyses and results. The remaining 13 articles yielded 22 *p*-values for entry in the *p*-curve analysis. However, 12 of these 22 *p*-values were statistically nonsignificant and so were automatically omitted from the analysis.[4] This resulted in a final total of 10 *p*-values for our *p*-curve analysis taken from seven articles. The *p*-curve generated is presented in Figure 1. A condensed version of our disclosure table that includes the relevant test statistics and *p*-values that can be incorporated into the *p*-curve app for reproduction of our analysis are included in Table 1. The full disclosure table is presented in our OSF repository and can be used to evaluate our analyses in greater detail (https://osf.io/mkynj/).

The *p*-curve analysis indicated that this set of studies does have evidential value. If either the left half of the *p*-curve is right skewed (at $\alpha = .05$) or both the full and half curve are right skewed (at $\alpha = .10$), then the *p*-curve suggests that these studies have good evidential value (Simonsohn et al., 2015). Both conditions were met (half: $z = -2.87$, $p = .002$; full: $z = -2.86$, $p = .002$).

Additionally, the *p*-curve app automatically reports a leave-one-out analysis, where the highest and lowest *p*-values are dropped from the analysis and the statistical significance of each test is recalculated. This determines whether the initial conclusion of the *p*-curve is robust to small changes in the test statistics included or dependent on just a few extreme *p*-values. For example, the smallest *p*-values are dropped one at a time, and the results of the tests after having removed the most extreme *p*-values are plotted (i.e.,

[4] The nature of the statistically nonsignificant results varied between articles. Some articles (e.g., Panero et al., 2016) did not predict that there would be an effect, and so their findings are in line with their predictions. In other cases, *p*-values near the .05 threshold were labeled as statistically significant, even when they were not. Finally, in one instance, a one-tailed test was used. Thus, the *p*-value calculated by the authors was manually halved, but the test statistic included in the *p*-curve analysis was associated with a *p*-value above .05 (i.e., not halved) and so was automatically omitted.

Figure 1
P-Curve of All Statistically Significant Results



Note. CI = confidence interval. *p*-curve app 4.06. See the online article for the color version of this figure.

the "drop K lowest original *p*-values" analysis). Sets of studies whose *p*-curve results remain consistent when extreme results are excluded are more trustworthy than those that change when only a few extreme results are omitted. The analysis is also specifically tuned to ensure that it does not overemphasize the effect of a single *p*-value (Simonsohn et al., 2015; osf.io/5yhqv/).

In the case of this *p*-curve, the exclusion of a single *p*-value (i.e., the smallest) does change the results: Neither the half *p*-curve nor the full *p*-curve is statistically significantly right skewed when the single lowest *p*-value is excluded from the analyses. This lack of statistical significance is maintained when the second, third, and fourth smallest *p*-values are also removed (Figure 2 plots the *p*-values of these *p*-curve tests [and is not itself a *p*-curve]). This analysis demonstrates that the initial *p*-curve results are contingent upon the inclusion of a single *p*-value, without which the *p*-curve no longer indicates good evidential value for this set of studies.

Although the *p*-curve is no longer significantly right skewed once the smallest *p*-value is dropped, the 33% power test is also not statistically significant. Thus, the *p*-curve does not conclude absent or inadequate evidential value. Rather, it indicates that the *p*-curve is not sufficiently powered to make a conclusion and that additional *p*-values are required. This is likely a reflection of the fact that over half of our *p*-values were dropped from the analysis for being greater than .05. This same result holds after one, two, three, or four of the smallest *p*-values are dropped from the analysis. In summary, removing the most extreme *p*-value results in a *p*-curve that is inconclusive, so the conclusion of the initial analysis is not robust.

The *p*-curve also estimates the average power in the set of studies included, and for these studies the average power is estimated to be 52% (90% CI [16, 81]). This suggests that many of the studies included in our sample had a very poor chance of identifying an effect, if such an effect exists. That said, this estimate itself is imprecise, as indicated by the wide confidence interval. Low statistical power may explain the mixed findings among these studies, the result of the leave-one-out analysis, and the fact that over half of the *p*-values originally identified as testing the central hypotheses were actually not statistically significant (and therefore could not be included in the *p*-curve).

Robustness *p*-Curve

As suggested by the creators of the *p*-curve, we also conducted a robustness *p*-curve, which involves running a *p*-curve analysis on an alternative set of test statistics (Simonsohn et al., 2014, 2015).[5] The selection of these alternative *p*-values was guided by the *p*-curve manual and is outlined in our preregistration (https://aspredicted.org/9es9g.pdf). In cases where the correct *p*-value to include was ambiguous or there was more than one appropriate option, we used the alternative choices in this robustness *p*-curve. Notes on these decisions can be found in the disclosure table (under "coding notes"). The robustness *p*-curve included 23 *p*-values from 16 articles. However, 15 of these results were automatically omitted from the *p*-curve as they were above .05, resulting in a total of eight *p*-values.

The results of this robustness *p*-curve are in line with those of the primary *p*-curve (see the full report on our OSF page: https://osf.io/mkynj/). The left half of the *p*-curve was statistically significantly right skewed ($z = -1.76$, $p = .039$), which indicates that the studies have good evidential value.[6] However, the *p*-curve also estimates that the statistical power of the studies included was very low, 13% (90% CI [5, 56]), with a narrower confidence interval than in the primary analysis. Furthermore, when the lowest *p*-value is excluded based on the leave-one-out analysis, the curve is no longer statistically significantly right skewed; this indicates that the initial conclusion of evidential value is not stable. Notably, this single *p*-value is responsible for the statistically significant right skew in both the primary and robustness *p*-curve. Once this *p*-value is dropped, the 33% power test becomes statistically significant, indicating that this set of results lacks evidential value. However, this result creeps above statistical significance after two or three of the smallest *p*-values are dropped. Thus, this conclusion should not be treated as definitive, but rather as a cause for concern. Much like with the main *p*-curve, the robustness *p*-curve's leave-one-out analysis indicates that more *p*-values from high-powered studies are needed to make a confident conclusion about the evidential value of these studies.

The results of this robustness *p*-curve are remarkably similar to those of the primary *p*-curve analysis: Both initially indicate that

[5] As indicated in our preregistration, we also intended to conduct a second robustness check, a *p*-curve analysis on studies that compared narrative fiction to expository fiction. However, this was deemed unnecessary as many of the studies that estimated this comparison did not obtain statistically significant results (e.g., Panero et al., 2016; Samur et al., 2018). Of those few studies that did obtain statistically significant results (i.e., Kidd & Castano, 2013), most appeared in a single article that was already subjected to a *p*-curve (van Kuijk et al., 2018). Other studies did not report the analytic details necessary to include their test statistics in the *p*-curve. Thus, conducting a *p*-curve on this particular comparison would offer very little novel insight.

[6] It is worth noting that the full *p*-curve was not statistically significantly right skewed ($z = -0.81$, $p = .210$), but this is not a necessary condition for concluding evidential value if the left half of the curve is right skewed at $\alpha = .05$.

Table 1
List of Test Statistics and Associated p-Values Included in the Analyses

Article	Test statistics	*p*-values	Robustness test statistic	Robustness *p*-values
Bal and Veltkamp (2013)	$F(1, 60) = 3.95$	$p = .051$ (NS)	$F(1, 60) = 3.95$	$p = .051$ (NS)
	$F(1, 86) = 4.56$	$p = .036$	$F(1, 86) = 4.56$	$p = .036$
Black and Barnes (2015)	$t(58) = 1.71$	$p = .093$ (NS)	$t(58) = 1.71$	$p = .093$ (NS)
Collins et al. (2017)	$t(19) = 0.582$	$p = .57$ (NS)	$t(19) = 0.582$	$p = .57$ (NS)
Gavaler and Johnson (2017)	$F(1, 193) = 3.81$	$p = .052$ (NS)	$F(1, 193) = 3.81$	$p = .052$ (NS)
Johnson et al. (2013)	$t(58) = 2.65$	$p = .01$	$t(56) = 1.15$	$p = .255$ (NS)
	$t(75) = 3.19$	$p = .002$	$t(74) = 1.77$	$p = .08$ (NS)
Kidd and Castano (2013)	$F(1, 82) = 6.40$	$p = .013$	$F(1, 82) = 6.40$	$p = .013$
	$t(74) = 1.93$	$p = .057$ (NS)	$t(74) = 1.90$	$p = .062$ (NS)
	$F(1, 65) = 4.97$	$p = .029$	$F(1, 65) = 4.97$	$p = .029$
	$F(1, 68) = 4.39$	$p = .04$	$F(1, 68) = 4.39$	$p = .04$
	$t(236) = 1.87$	$p = .06$ (NS)	$t(236) = 2.34$	$p = .02$
Kidd and Castano (2019)	$F(1, 342) = 4.91$	$p = .027$	$F(1, 341) = 3.90$	$p = .049$
Kidd et al. (2016)	$t(210) = 2.28$	$p = .0236$	$t(210) = 1.96$	$p = .051$ (NS)
Koopman et al. (2012)	$\chi^2(1) = 3.84$	$p = .050$ (NS)	$\chi^2(1) = 3.84$	$p = .05$ (NS)
Panero et al. (2016)	$F(1, 167) = 0.08$	$p = .775$ (NS)	$F(1, 191) = 0.83$	$p = .363$ (NS)
			$F(1, 53.6) = 0.08$	$p = .775$ (NS)
Pino and Mazza (2016)	$t(139) = 6.24$	$p < .00001$	$t(145) = 4.97$	$p < .00001$
Samur et al. (2018)	$F(1, 153) = 0.56$	$p = .45$ (NS)	$F(1, 153) = 0.56$	$p = .45$ (NS)
	$F(1, 153) = 3.31$	$p = .07$ (NS)	$F(1, 151) = 0.01$	$p = .92$ (NS)
	$F(3, 319) = 0.29$	$p = .84$ (NS)	$F(3, 319) = 0.29$	$p = .84$ (NS)
	$F(1, 128) = 0.47$	$p = .49$ (NS)	$F(1, 125) = 0.76$	$p = .39$ (NS)
van Kuijk et al. (2018)	$F(1, 381) = 12.275$	$p = .00051$	$F(1, 389) = 7.566$	$p = .006$

Note. NS = not significant.

the studies have good evidential value but that this conclusion is also mercurial and contingent on the inclusion of the same smallest *p*-value. In this way, the lack of robustness discussed earlier is itself robust as it is consistent across two iterations of the *p*-curve. Both *p*-curves also estimated that the sampled studies had low statistical power and were thus unlikely to correctly identify a true effect if one exists.

Discussion

The goal of this study was to investigate the experiments assessing the causal effect of reading literary fiction on social cognition. Findings in this literature have been mixed, and we assessed the possibility that results from this literature may have been influenced by selective reporting, or *p*-hacking, by using a *p*-curve analysis (Simonsohn et al., 2014). We conducted a *p*-curve analysis on all published experimental estimates of the effect of reading literary fiction on social cognition that were amenable to inclusion in a *p*-curve. The initial results of this analysis indicated that the studies do not lack evidential value and are thus unlikely to have been meaningfully impacted by selective reporting. However, these results were fragile and contingent upon the inclusion of a single *p*-value. Further, this fragility was also observed in the robustness *p*-curve that used alternative *p*-values from the same set of studies. This suggests that we should be cautious in interpreting the initial conclusion of this *p*-curve analysis. Incidentally, it also highlights the dangers of making binary decisions based on whether a *p*-value is above or below a given threshold.

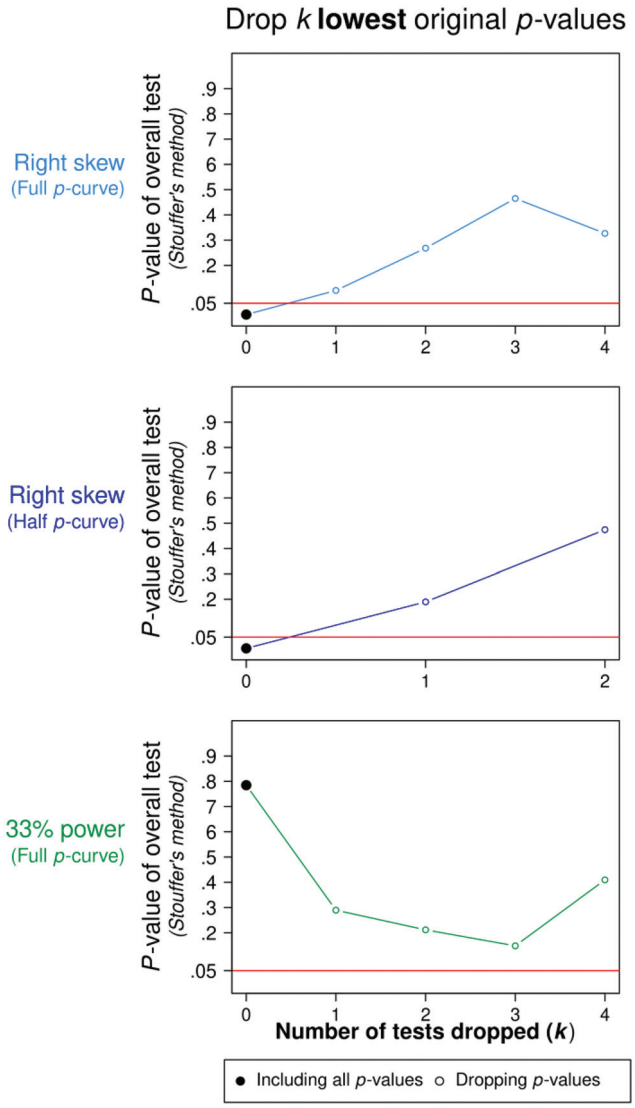
Although the results of these analyses are not definitive (i.e., it is unclear whether these studies have good or inadequate evidential value), this lack of conclusiveness is itself a form of conclusion: We cannot be confident in the current literature on the causal effect of reading fiction on social cognition. We base this conclusion not just on the results of the *p*-curves but also on the fact that over half of the studies we collected did not report a statistically significant main effect (12 of 22). Further, some of the studies that did claim to find such an effect actually had statistically nonsignificant results. Although the *p*-curve does not include statistically nonsignificant results in its analysis, they are informative when considering the uniformity of evidence in this experimental literature more broadly. Additionally, the *p*-curve analyses also estimated that the studies that did find statistically significant results had levels of statistical power much lower than what is typically desired and were thus unlikely to correctly identify a true effect.

Some recent findings do present reasons for optimism in this literature, however. For example, van Kuijk and colleagues' (2018) high-powered replication of the original Kidd and Castano (2013) study suggests that a short exposure to literary fiction may indeed cause a small but detectable improvement in some social cognitive abilities. However, the results of our *p*-curve analyses indicate that additional adequately powered studies are needed. At present, this literature remains quite mixed and does not provide strong evidence for the effect. It remains unclear whether past findings have been the result of selective reporting and whether future replication attempts are likely to be successful. It also seems unlikely that past work has had sufficient statistical power to detect true effects of this magnitude.

Another possible explanation for these mixed results is the mismatch between experimental studies and the theoretical accounts of how literary fiction is likely to affect social cognition. Reading fiction likely improves social cognitive abilities over long stretches of time, through the cumulative effect of frequent, prolonged, and volitional exposures to literary fiction (Mar, 2018a). This form of exposure cannot be captured in experimental designs, which typically present brief narratives in a single nonvolitional instance. Kidd and Castano (2013, 2017, 2019) proposed that a single exposure to

Figure 2

Examining Stability of the Results by Omitting the Lowest p-Values Using the Leave-One-Out Analyses Included in the p-Curve Report



Note. Both the full and half *p*-curve fall above threshold for statistical significance after the single lowest value is omitted. *p*-curve app 4.06. See the online article for the color version of this figure.

literary fiction can improve performance on theory of mind tasks by priming the social cognitive processes associated with understanding others. They further state that this effect is unique to literary fiction, which requires readers to override automatic script-based thinking in favor of more effortful theory of mind processes. This is placed in contrast to popular fiction, which features more formulaic and stereotypical content and therefore does not require readers to engage in effortful social cognition. Many priming effects, however, have proven to be sensitive and difficult to replicate (Cesario, 2014), which may be why this immediate effect of fiction on social cognition appears inconsistently.

It is also worth noting that the *p*-curve is not without its limitations. For example, some have argued that the *p*-curve underperforms in effect size estimation (compared to other similar methods) when there is considerable heterogeneity in the effect sizes analyzed (McShane et al., 2016; cf. but disputed by the *p*-curve authors; Simmons et al., 2018). Other attempts to evaluate the *p*-curve have shown that it performs similarly to other techniques with comparable goals, but its effect size estimation suffers when there is no true effect (Carter et al., 2019). Note that our aim was not to estimate the magnitude of any effect, but rather to assess the evidential value for this body of literature (i.e., the primary aim of the *p*-curve method). Readers who prefer other methods for evaluating evidential value or estimating effect size are encouraged to take advantage of our publicly posted data in order to do so.

The present study found that the current experimental literature on the causal effect of reading fiction on social cognition has only tenuous evidential value. Although our results do not indicate that a true causal effect does not exist, they do encourage caution when interpreting past demonstrations of this effect. A *p*-curve analysis was not able to definitively indicate whether this work has been meaningfully impacted by selective reporting or *p*-hacking. However, these results do highlight several potential problems with the literature that will need to be addressed moving forward: low statistical power, potential selective reporting, and unclear theoretical accounts for the effect. Fortunately, these problems are solvable. For instance, some work in this area has already demonstrated the benefits of adequately powered studies (e.g., Samur et al., 2018; van Kuijk et al., 2018). Further, the issue of selective reporting can largely be addressed by thorough preregistration and the transparent reporting of methodology and analyses. Incidentally, preregistered experiments on this topic have failed to yield consistent evidence in support of an immediate influence on social cognition following a narrative presentation (Kidd & Castano, 2019; Samur et al., 2018). Finally, theoretical accounts of this effect should reflect the differences between short experimental exposures and long-term repeated volitional exposure to stories. This may also require the consideration of alternative theoretical accounts, such as the role of priming in this putative effect.

Context of the Research

Our lab has been studying the association between stories and social cognition for over 15 years, beginning with the doctoral dissertation of the senior author (Raymond A. Mar). The question of whether exposure to narratives helps to foster social abilities has now been examined using a wide range of methodologies by labs all over the world. This includes developmental work with children on how parent–child reading fosters theory of mind, as well as neuroscience investigations into whether the brain networks supporting story comprehension and social cognition overlap. Correlational studies of lifetime exposure to narrative texts have also explored the boundary conditions of this association, examining which genres have the strongest associations with mentalizing. Theoretically, we have developed a research framework to guide research on this topic, the SPaCEN framework, which specifies two pathways through which stories might foster social cognition. This *p*-curve analysis of past experimental research is a natural extension of this framework and the ongoing work in our lab.

References

Articles indicated with an asterisk were included in the *p*-curve analysis.

Adrian, J. E., Clemente, R. A., Villanueva, L., & Rieffe, C. (2005). Parent-child picture-book reading, mothers' mental state language and children's theory of mind. *Journal of Child Language*, *32*(3), 673–686. https://doi.org/10.1017/S0305000905006963

*Bal, P. M., & Veltkamp, M. (2013). How does fiction reading influence empathy? An experimental investigation on the role of emotional transportation. *PLoS ONE*, *8*(1), Article e55341. https://doi.org/10.1371/journal.pone.0055341

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715

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

Blair, R. J., & Cipolotti, L. (2000). Impaired social response reversal. A case of 'acquired sociopathy.' *Brain: A Journal of Neurology*, *123*(6), 1122–1141. https://doi.org/10.1093/brain/123.6.1122

Boyd, B. (2009). *On the origin of stories: Evolution, cognition, and fiction*. Harvard University Press.

Carruthers, P., & Smith, P. K. (Eds.). (1996). *Theories of theories of mind*. Cambridge University Press. https://doi.org/10.1017/CBO9780511597985

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

*Collins, K. L., Zweber, A., & Irwin, A. N. (2017). Impact of a fictional reading intervention on empathy development in student pharmacists. *Currents in Pharmacy Teaching & Learning*, *9*(3), 498–503. https://doi.org/10.1016/j.cptl.2016.12.003

Confucius. (1861). *The analects* (J. Legge, Trans.). http://classics.mit.edu (Original work published c. 500 BCE)

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*(5), 725–730. https://doi.org/10.1037/a0013283

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *Catalog of Selected Documents in Psychology*, *10*, 85.

Dodell-Feder, D., & Tamir, D. I. (2018). Fiction reading has a small positive impact on social cognition: A meta-analysis. *Journal of Experimental Psychology: General*, *147*(11), 1713–1727. https://doi.org/10.1037/xge0000395

*Djikic, M., Oatley, K., & Moldoveanu, M. C. (2013). Reading other minds: Effects of literature on empathy. *Scientific Study of Literature*, *3*(1), 28–47. https://doi.org/10.1075/ssol.3.1.06dji

Fong, K., Mullin, J. B., & Mar, R. A. (2015). How exposure to literary genres relates to attitudes toward gender roles and sexual behavior. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(3), 274–285. https://doi.org/10.1037/a0038864

*Gavaler, C., & Johnson, D. (2017). The genre effect: A science fiction (vs. realism) manipulation decreases inference effort, reading comprehension, and perceptions of literary merit. *Scientific Study of Literature*, *7*(1), 79–108. https://doi.org/10.1075/ssol.7.1.04gav

Gerrig, R. J. (1993). *Experiencing narrative worlds*. Yale University Press. https://doi.org/10.12987/9780300159240

Higgins, J. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). (Eds.). *Cochrane handbook for systematic reviews of interventions* (Version 6.0). www.training.cochrane.org/handbook

Hogan, P. C. (2003). *The mind and its stories: Narrative universals and human emotion*. Cambridge University Press. https://doi.org/10.1017/CBO9780511499951

*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*(5), 578–598. https://doi.org/10.1521/soco.2013.31.5.578

*Kevane, M. (2020). Reading fiction and economic preferences of rural youth in Burkina Faso. *Economic Development and Cultural Change*, *68*(3), 1041–1079. https://doi.org/10.1086/701704

*Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*(6156), 377–380. https://doi.org/10.1126/science.1239918

Kidd, D. C., & Castano, E. (2017). Panero et al. (2016): Failure to replicate methods caused the failure to replicate results. *Journal of Personality and Social Psychology*, *112*(3), e1–e4. https://doi.org/10.1037/pspa0000072

*Kidd, D., & Castano, E. (2019). Reading literary fiction and theory of mind: Three preregistered replications and extensions of Kidd and Castano (2013). *Social Psychological & Personality Science*, *10*(4), 522–531. https://doi.org/10.1177/1948550618775410

*Kidd, D., Ongis, M., & Castano, E. (2016). On literary fiction and its effects on theory of mind. *Scientific Study of Literature*, *6*(1), 42–58. https://doi.org/10.1075/ssol.6.1.04kid

*Koopman, E. M., Hilscher, M., & Cupchik, G. C. (2012). Reader responses to literary depictions of rape. *Psychology of Aesthetics, Creativity, and the Arts*, *6*(1), 66–73. https://doi.org/10.1037/a0024153

*Kuzmičová, A., Mangen, A., Støle, H., & Begnum, A. C. (2017). Literature and readers' empathy: A qualitative text manipulation study. *Language and Literature*, *26*(2), 137–152. https://doi.org/10.1177/0963947017704729

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine*, *6*(7), Article e1000100. https://doi.org/10.1371/journal.pmed.1000100

Mar, R. A. (2018a). Evaluating whether stories can promote social cognition: Introducing the social processes and content entrained by narrative (SPaCEN) framework. *Discourse Processes*, *55*(5–6), 454–479. https://doi.org/10.1080/0163853X.2018.1448209

Mar, R. A. (2018b). Stories and the promotion of social cognition. *Current Directions in Psychological Science*, *27*(4), 257–262. https://doi.org/10.1177/0963721417749654

Mar, R. A., & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science*, *3*(3), 173–192. https://doi.org/10.1111/j.1745-6924.2008.00073.x

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*(5), 694–712. https://doi.org/10.1016/j.jrp.2005.08.002

Mar, R. A., Tackett, J. L., & Moore, C. (2010). Exposure to media and theory-of-mind development in preschoolers. *Cognitive Development*, *25*(1), 69–78. https://doi.org/10.1016/j.cogdev.2009.11.002

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

Mertens, G., & Krypotos, A.-M. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, *59*(1), 338–352. https://doi.org/10.5334/pb.493

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 109–120. https://doi.org/10.1037/aca0000089

*Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2016). Does reading a single passage of literary fiction really improve theory of mind? An attempt at replication. *Journal of Personality and Social Psychology*, *111*(5), e46–e54. https://doi.org/10.1037/pspa0000064

Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2017). No support for the claim that literary fiction uniquely and immediately improves theory of mind: A reply to Kidd and Castano's commentary on Panero et al. (2016). *Journal of Personality and Social Psychology*, *112*(3), e5–e8. https://doi.org/10.1037/pspa0000079

*Pino, M. C., & Mazza, M. (2016). The use of "literary fiction" to promote mentalizing ability. *PLoS ONE*, *11*(8), Article e0160254. https://doi.org/10.1371/journal.pone.0160254

Preston, S. D., & de Waal, F. B. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, *25*(1), 1–20. https://doi.org/10.1017/S0140525X02000018

Quinlan, J., Padgett, J. K., Khajehnassiri, A., & Mar, R. A. (2021). *Test statistics for a p-curve on experiments testing whether a brief exposure to literary fiction cause increases in social ability* [Data set]. osf.io/mkynj/

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow & A. Collins (Eds.), *Representation and understanding: Studies in cognitive science* (pp. 2–34). Academic Press. https://doi.org/10.1016/B978-0-12-108550-6.50013-6

*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*(1), 130–144. https://doi.org/10.1080/02699931.2017.1279591

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018, January 8). *P-curve handles heterogeneity just fine*. http://datacolada.org/67

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*, *28*(5), 687–693. https://doi.org/10.1177/0956797616658563

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Stein, N. L., & Glenn, C. G. (1975). *An analysis of story comprehension in elementary school children: A test of a schema*. https://eric.ed.gov/?id=ED121474

*van Kuijk, I., Verkoeijen, P., Dijkstra, K., & Zwaan, R. A. (2018). The effect of reading a short passage of literary fiction on theory of mind: A replication of Kidd and Castano (2013). *Collabra: Psychology*, *4*(1), Article 7. https://doi.org/10.1525/collabra.117

Weston, S. J., Ritchie, S. J., Rohrer, J. M., & Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, *2*(3), 214–227. https://doi.org/10.1177/2515245919848684

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

Received June 14, 2021
Revision received April 15, 2022
Accepted August 4, 2022 ■