Journal of Personality and Social Psychology
2017, Vol. 112, No. 3, e5–e8

0022-3514/17/$12.00 http://dx.doi.org/10.1037/pspa0000079

REPLY

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)

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

Kidd and Castano (in press) critique our failure to replicate Kidd and Castano (2013) on 3 grounds: failure to exclude people who did not read the texts, failure of random assignment, and failure to exclude people who did not take the Author Recognition Test (ART). This response addresses each of these critiques. Most importantly, we note that even when Kidd and Castano reanalyzed our data in the way that they argue is most appropriate, they still failed to replicate the pattern of results reported in their original study. We thus reaffirm that our replication of Kidd and Castano (2013) found no evidence that literary fiction uniquely and immediately improves theory of mind. Our objective remains not to prove that reading literary fiction does not benefit social cognition, but to call for in-depth research addressing the difficulties in measuring any potential effect and to note the need to temper claims accordingly.

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

Kidd and Castano (2013) reported that reading a short piece of literary fiction (compared to reading popular fiction, reading nonfiction, or not reading at all) immediately and significantly boosts theory of mind, as measured by performance on the Reading the Mind in the Eyes Test (RMET; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). When a surprising and potentially important finding such as this is published, replication is called for. In addition, the original report did not treat individual texts as a random effect, which leaves open an increased potential for Type I errors (Judd, Westfall, & Kenny, 2012). Three labs independently carried out replication studies and later combined their data into a single analysis. This replication study found no advantage in RMET scores for literary fiction in comparison with any of the other conditions (Panero et al., 2016).

Kidd and Castano (in press) make three criticisms of our replication effort: failure to exclude people who did not read the texts, failure of random assignment, and failure to exclude people who did not take the ART (Acheson, Wells, & MacDonald, 2008; Stanovich & West, 1989). Before responding specifically to these criticisms, it is crucial to note that even accepting all of Kidd and Castano's (in press) adjustments to our data set, their reanalysis does not replicate their 2013 finding. Although their new analyses do show that performance in the literary fiction condition is higher than in the popular fiction condition, they find that performance in the literary fiction condition is no better than in the expository nonfiction condition.

These findings not only cast doubt on their claim that literary fiction has a unique ability to improve theory of mind, but also raise questions about the theoretical underpinning of the prediction that literary fiction should improve theory of mind. Kidd and Castano argued that literary fiction achieves this effect

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.

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

because it invites readers to adopt a writerly perspective (e.g., Kidd & Castano, 2013) or because it features "round" rather than "flat" characters (e.g., Kidd & Castano, 2016). Given that no difference was found between the literary fiction condition and the expository (i.e., nonsocial) nonfiction condition, *neither* of these theories is consistent with the data, even after substantial reanalysis. We return to this key point in the final section of our reply.

Reading Time Exclusions

All of these studies involve assigning participants to read texts and measuring subsequent performance on the RMET. Kidd and Castano (in press) correctly note that it is important to exclude participants whose reading times are so fast that they cannot have had adequate exposure to their assigned text. In Experiments 3, 4, and 5, Kidd and Castano (2013) excluded participants who spent fewer than 30 s on any page of text, but they did not have a systematic exclusion criterion for fast reading in Experiments 1 and 2. They excluded one participant whose total reading time was 0 s and one whose reading time was 1.13 s per page in Experiments 1 and 2, respectively. In addition, in all five experiments they excluded participants whose reading speeds were greater than 3.5 *SD* above the mean (hence too slow). We adopted this 3.5 *SD* cutoff in our replication as well, for reading times 3.5 *SD* either above or below the mean. However, none of our participants had scores below the 3.5 *SD* cutoff for reading time.

In their reanalysis of our data, Kidd and Castano argue that this criterion is insufficient and instead use a formula for reading time exclusion based on a paper by Bell and Perfetti (1994). This study had tiny sample sizes for skilled readers (10 per cell), which casts doubt on the validity of their suggested formula. Although we agree that it is important to verify that participants have read their assigned texts, there is a potential downside to using reading time as a manipulation check. In addition to reflecting *whether* participants have read a given text, reading time is also likely to reflect *how* they read that text. Zwaan (1991) found that those who read a passage they believe is fiction read more slowly than when they believe the same passage is nonfiction. Furthermore, it has been argued that specifically literary fiction, but not popular fiction, evokes a kind of mental stillness during which readers stop and reflect as they read (Koopman & Hakemulder, 2015). Similarly, readers of popular fiction—particularly those who are familiar with a given genre—may use their knowledge of the genre to facilitate reading and understanding, which could potentially allow them to read more quickly (see discussion of "reading in a system" in Carroll, 1994). Indeed, popular fiction may sometimes be seen as appealing *because* it can be read quickly.

Thus, although it is crucial to ensure that participants have read the material they have been given, applying too strict a criterion for reading time may systematically exclude participants based on their familiarity with or their motivation to read the kind of material they have been assigned. The reading speed chosen by Kidd and Castano (in press) in their reanalysis of our nonreplication appears to be problematic in this way, as it resulted in participants being excluded disproportionately from the popular fiction condition. Given that participants would be *expected* to read popular fiction more quickly than literary fiction, it seems likely that the criterion they applied is not merely excluding participants who were "not exposed" to the experimental manipulation, but also individuals who *did* read the assigned material, but were able to read more quickly than individuals in the nonfiction or literary fiction conditions.

Nonetheless, because we agree that it is important to only include participants who spent adequate time reading their assigned text, we reanalyzed our full data set with an eye toward setting a criterion for reading time that would ensure that participants had been exposed to the source material, while simultaneously allowing for variation in reading times between conditions (raw data available at osf.io/83nv2/). Thus, we began by excluding all participants whose reading time was ≥600 words per minute (WPM; hence too fast). We chose this cutoff based on literature reviewed by Lewandowski, Codding, Kleinmann, and Tucker (2003), who cite average WPM values for college students from 136 to 400 WPM. Six hundred WPM is an arbitrary cutoff, but a reasonable one that is consistent with how skilled readers may approach reading fiction.

We obtained a WPM reading time by summing the times spent on the individual pages for a text and then dividing the total number of words in each text by the total number of minutes to read all the pages. We then excluded all participants (109) whose WPM rates were greater than 600. The median of the remaining sample was 262 WPM, and the mean was 287 (SD = 110), in line with the literature cited by Lewandowski et al. (2003).

The pattern of means from this reduced sample again fails to replicate the original study (see Table 1 for mean scores by lab and by condition). The results also remain unchanged when we excluded those who read over 500 WPM (reading too fast), and when we excluded those who read fewer than 100 WPM (reading too slowly).

Given the problems outlined above with using reading time as a manipulation check, as well as the challenge in determining where the (somewhat arbitrary) line should be drawn to distinguish between participants who have and have not read the given material, we recommend that future research on this topic employ a reading comprehension manipulation check (e.g., Black & Barnes, 2015; Pino & Mazza, 2016).

Random Assignment

Kidd and Castano (in press) state that because of unequal sample sizes across conditions, Labs 1 and 2 from Panero et al. (2016) failed to randomize properly. This is a justified critique given a lack of clarity describing the method of Research Group 1. We did assign reading texts randomly. However, Research Group 1 used a control (no-reading) group that, although sampled from the same population (Mechanical Turk; MTurk), was not part of the random assignment to reading conditions for Experiment 5. Specifically, Research Group 1's Experiment 5 appeared on MTurk as two postings, one for the no-reading condition and one that included the two other conditions: literary and popular fiction. The two postings were identical in their description and compensation, and thus there is no reason to believe that the control participants differed in any way from those in the other conditions. Crucially, for those participants who enrolled in the Experiment 5 posting containing the literary fiction and popular fiction conditions, assignment of participants to text was fully random. It is also important to note that Research Group 1 included checks so that

participants could not sign up for more than one experiment, hence no participant was part of the data set twice. And even if we exclude our third analysis, which was the only one to include the no-reading control, our comparisons between literary fiction and nonfiction and between literary fiction and popular fiction both failed to replicate the original finding.

With respect to unequal sample sizes, Research Group 1 capped each MTurk posting based on the sample sizes originally recruited for each experiment in Kidd and Castano (2013). The only cell that deviates from this is Experiment 5—Reading. Due to funding constraints, Research Group 1 had to stop data collection early, resulting in fewer participants than originally planned in the reading conditions for Experiment 5.

Kidd and Castano (in press) correctly point out that Research Group 2 had over twice as many participants in the literary fiction condition than in the nonfiction and no-reading conditions. This occurred because, as they note in their Footnote 3, literary fiction participants were randomly assigned to read the same text described as either fiction or nonfiction. As explained in Panero et al. (2016), because no RMET differences were found across these two conditions, the data were collapsed into one literary fiction condition. In addition, there was a slight error in one of the randomization counters in Qualtrics, which led to more participants being enrolled in this condition ($n = 41$). While this does mean that participant numbers were not equally matched across conditions, we disagree that this "compromises the internal validity of the experiment." More importantly, our data fail to replicate Kidd and Castano (2013) even if we exclude all of the data from Research Group 2. Thus this concern does not affect our primary conclusions.

ART Scores

The ART presents participants with a list of names and asks them to select only the real authors from this list. Half of the names are real authors and half are foils, so scores on the ART are calculated by subtracting the number of foils selected from the number of authors correctly selected (Acheson et al., 2008; Stanovich & West, 1989).

Kidd and Castano (in press) argue that participants who did not select any items on the ART should be excluded. In their comment (p. 6) they stated that they excluded eight participants "who selected no items, foils or authors, on the ART." We had assumed, based on inspections of the data, that these people had simply not recognized any names. However, except in the case of Research Group 3, which included a manipulation check, we have no way to be certain that those who selected no names actually engaged with the task. We agree that it makes sense to exclude participants who did not complete a key covariate in these circumstances, although we found 21 such participants, rather than the eight identified by Kidd and Castano (in press). This discrepancy is most likely due to the fact that they first excluded participants based on considerations of reading time.

To investigate whether excluding these 21 cases would affect our results, we performed a series of post hoc analyses that replicated the three comparisons reported in our original paper (Panero et al., 2016).[1] This reanalysis did not shift the results across the $p < .05$ significance level for either the literary fiction versus nonfiction comparison ($N = 1$ case excluded; $F(1, 166) = 0.13$, $p = .720$) or for the literary fiction versus popular fiction comparison ($N = 17$ cases excluded; $F(1, 199) = 0.14$, $p = .707$). Although Kidd and Castano (in press) did not include the no-reading control in their reanalysis of our data, we included that comparison here. Strikingly, the removal of cases with zero ART scores from this condition ($N = 4$)[2] did alter the outcome of our results, with literary fiction conveying a benefit over the no-reading control, $F(1, 54.6) = 4.11$, $p = .048$.[3] Although the effect size was very small ($d = 0.01$) and the p value was high for a post hoc test, this raises points for consideration in future research. First, because including people who do not recognize any names or select foils may be important, we suggest that ART-type measures should include an "I do not recognize any of these names as authors" option. Second, considering the lack of evidence that literary fiction reliably improves theory of mind compared with nonfiction and (depending on exclusion criteria) popular fiction, the results of this analysis suggest that no-reading control conditions should be compared with reading other narratives besides literary fiction.

Kidd and Castano (in press) Fails to Replicate Kidd and Castano (2013)

Even after all of the exclusions carried out on our data, Kidd and Castano (in press) fail to replicate their original pattern of findings, with RMET scores higher in the literary fiction condition than in all other conditions (popular fiction, nonfiction, no-reading) and with no difference between popular fiction and no-reading. Although they found that individuals assigned to read literary fiction performed better than those assigned to read popular fiction on the RMET, the literary fiction group did no better than the nonfiction group. The nonfiction stimuli used in these experiments are not only nonfiction, but also largely *nonsocial*. Thus, it is difficult to interpret this pattern of results using any theoretical framework about the power of literary fiction to engage our capacity for theory of mind (e.g., Kidd & Castano, 2013, 2016). Moreover, nonfiction readers actually performed significantly better than popular fiction readers on the RMET. Given their original claim that literary fiction boosts theory of mind, this particular result provides strong evidence against this claim; if anything, the reanalysis of the data might suggest that popular fiction impedes theory of mind compared with *both* literary fiction and nonfiction. At the very least, the reanalysis presented by Kidd and Castano (in press) confirms that literary fiction did not uniquely increase theory of mind performance.

Moreover, the only statistically significant effect found by Kidd and Castano's (in press) reanalysis in the hypothesized direction (literary fiction > popular fiction), had a p value of .044. Our own reanalysis, described in the previous section, revealed one significant effect on a comparison not included in Kidd and Castano (in press): RMET scores in the literary fiction condition were higher

[1] In all three analyses, we used SAS (9.4) PROC MIXED, restricted maximum likelihood estimation, Kenward-Roger degrees of freedom, with text as a random variable nested within experiment and condition. Condition, ART scores, and their interaction predicted RMET scores. Skewed variables were transformed to meet assumptions.

[2] One case overlapped with the literary vs. popular fiction comparison.

[3] The effect of ART and its interaction with condition were statistically significant, mirroring the analysis reported in Panero et al. (2016).

than in the no-reading control condition, $p = .048$. However, using an alpha value of $p < .05$ is problematic with multiple comparisons and replications (Lindsay, 2015; Murayama, Pekrum, & Fiedler, 2014; Stangor & Lemay, 2016). Given that the p values for the literary versus popular fiction comparisons in the original paper (Kidd & Castano, 2013) were all $p = .04$, this effect may be particularly tenuous. The Open Science Collaboration (2015) found that only 18% of studies with $.04 < ps < .05$ replicated. In part, this may be due to small sample sizes used in replication attempts (Maxwell, Lau, & Howard, 2015), but it serves as a warning and a call for replication efforts such as Panero et al. (2016) provide. As such, rather than understanding Kidd and Castano's (in press) and our own reanalyses as proof that our replication was poor, we take them as further evidence that the effect is likely to be ephemeral.

In discussing the lack of replication of literary fiction providing an RMET advantage over nonfiction, Kidd and Castano (in press) refer to work by Black and Barnes (2015), where a within-subjects design found that participants performed better on the RMET after reading literary fiction than nonfiction. We believe that it is worth noting that the authors of that study coauthored the Panero et al. (2016) nonreplication: After having obtained the effect using a within-subjects design, they were unable to obtain it between-subjects. Thus, the purpose of publishing the nonreplication was, in part, to acknowledge this inconsistency and use it to motivate further research that examines why this effect may be fragile, particularly in between-subjects, rather than within-subjects, designs.

Although we disagree with Kidd and Castano about the specifics of this dataset, our shared goal in this research project is to discover how exposure to fiction affects people's lives. We thus call on researchers to engage in more nuanced analyses of how stories might have an effect, and under what circumstances, and for whom.

References

Acheson, D. J., Wellu, 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

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

Bell, L. C., & Perfetti, C. A. (1994). Reading skill: Some adult comparisons. *Journal of Educational Psychology, 86,* 244–255. http://dx.doi.org/10.1037/0022-0663.86.2.244

Black, J. E., & 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

Carroll, N. (1994). The paradox of junk fiction. *Philosophy and Literature, 18,* 225–241. http://dx.doi.org/10.1353/phl.1994.0054

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

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

Kidd, D., & Castano, E. (2016). Different stories: How levels of familiarity with literary and genre fiction relate to mentalizing. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. http://dx.doi.org/10.1037/aca0000069

Kidd, D. C., & Castano, E. (in press). Panero et al. (2016): Failure to replicate methods caused the failure to replicate results. *Journal of Personality and Social Psychology*.

Koopman, E. M. E., & Hakemulder, F. (2015). Effects of literature on empathy and self-reflection: A theoretical-empirical framework. *Journal of Literary Theory, 9,* 79–111.

Lewandowski, L. J., Codding, R. S., Kleinmann, A. E., & Tucker, K. L. (2003). Assessment of reading rate in postsecondary students. *Journal of Psychoeducational Assessment, 21,* 134–144. http://dx.doi.org/10.1177/073428290302100202

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26,* 1827–1832. http://dx.doi.org/10.1177/0956797615616374

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering From a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70,* 487–498. http://dx.doi.org/10.1037/a0039400

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18,* 107–118. http://dx.doi.org/10.1177/1088868313496330

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. http://dx.doi.org/10.1126/science.aac4716

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. http://dx.doi.org/10.1037/pspa0000064

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

Stangor, C., & Lemay, E. P. (2016). Introduction to the Special Issue on Methodological Rigor and Replicability. *Journal of Experimental Social Psychology, 66,* 1–3. http://dx.doi.org/10.1016/j.jesp.2016.02.006

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

Zwaan, R. A. (1991). Some parameters of literary and news comprehension: Effects of discourse-type perspectives on reading rate and surface structure representation. *Poetics, 20,* 139–156.

Received December 12, 2016
Revision received January 9, 2017
Accepted January 10, 2017 ■