Reading Others as Individuals: A Field Experiment on Fiction and Prosociality*

Juan Ferro José-Alberto Guerra

March 2026

Abstract

What drives prosocial behavior and how it can be shaped through education interventions? This paper examines whether reading literary fiction can causally influence prosociality. In a pre-registered field experiment, university students were randomly assigned to one of two four-week extracurricular courses: a fiction-reading curriculum or a placebo course focused on behavioral science, with equivalent intensity and reading content. After the intervention, participants took part in experimental sessions featuring incentivized economic games and psychological scales. We find evidence that the fiction course increases prosocial behavior, primarily through higher altruism and greater trustworthiness. The effects appear to reflect shifts in individual preferences rather than updates in social norms. While empathy does not change, we find suggestive evidence that fiction increases transportation and attention to what makes others unique—consistent with a social cognition mechanism centered on valuing individuation rather than self–other overlap or morality. These findings contribute to the growing literature on how narrative experiences shape social behavior and offer paths for future interventions to increase prosociality without social norms change.

JEL Classification Codes: C93, D91, Z11, I24, D63
Keywords: prosociality, fiction-reading, individuation, empathy, theory of mind.

*Financial support from the Teaching and Researching Equitable Economics from the South (TREES) initiative at Universidad de Los Andes is gratefully acknowledged. We thank Kelly Daniela Cadena and María Alejandra Páez for their excellent research assistance. The paper has benefited from comments by Andrés Álvarez, Nava Ashraf, Jorge Luis Baquero, José Luis Falconi, James Robinson, and audiences at the BEEC, Universidad Javeriana, and TREES workshop at Universidad de Los Andes. The filed experiment was approved by Universidad de los Andes' and Department of Economics' IRB by Acta No 88 de 2025.

1 Introduction

"It is for this reason that art in general, literature especially (...) is not exactly favored by the (...) masters of the masses (...) into the little zeros with which [they] tend to operate, art introduces a "period, period, comma, and a minus", transforming each zero into a tiny human, albeit not always pretty, face."

Nobel Lecture (1987), Joseph Brodsky

The notion that reading literary fiction is associated with desirable individual behavior is widespread in moral philosophy (Brodsky, 1987; Keen, 2007; Nussbaum, 1990; Zunshine, 2006). Yet this relationship may be non-causal: people who choose to read fiction could differ systematically from those who do not, generating a pure selection effect (Dubourg & Baumard, 2023). This article uses tools from experimental economics and behavioral science to address this concern. We ask whether participation in a guided fiction-reading curriculum causally affects prosociality, understood as behaviors and judgments that favor others' welfare, and we use a combination of incentivized behavioral measures and psychological scales to shed light on the mechanisms.

A large empirical literature documents correlations between fiction reading and prosocial outcomes, but rigorous causal evidence remains limited. Some meta-analysis suggest a small positive effect of fiction reading on social cognition—an umbrella that encompasses theory of mind and empathy (Dodell-Fedder & Tamir, 2018) and short-run gains in theory of mind after reading literary fiction (Kidd & Castano, 2013). At the same time, recent randomized studies that extend the intervention horizon and broaden outcomes have reported null effects on social outcomes, highlighting the need for designs that both strengthen identification and clarify mechanisms (Dodell-Feder et al., 2025). Our study contributes to this debate with a randomized field intervention implemented as an extracurricular course, and with outcomes that emphasize revealed-preference behavior rather than survey responses alone. Specifically, we use random assignment to a fiction-reading course (treatment) or to a parallel non-fiction course on behavioral science (placebo). Both courses consist of four 90-minute in-person sessions delivered over four weeks by instructors who are experts in their fields. The courses are designed to equalize attention and participation while isolating the role of literary narrative.

A common hypothesis is that literary fiction can shape prosocial attitudes and behaviors by strengthening social cognition—most prominently empathy ("can I feel what you feel?"), narrative immersion or transportation ("can I be pulled into the story?"), and theory of mind ("can I imagine what you think?") (Dodell-Fedder & Tamir, 2018; Green & Brock, 2000a; Kidd & Castano, 2013). We adopt this broad view, but we also emphasize an additional mechanism that is less commonly measured in experimental work on fiction reading: a shift in how individuals represent *the other* in the first place. Specifically, engaging with literary narratives may increase the tendency to perceive others as distinctive, individuated persons rather than as interchangeable members of a category.

This uniqueness/individuation channel ("can I attend to what makes you unique?") is conceptually related to classic work on stereotyping and impression formation, where individuated processing is contrasted with categorical processing (Fiske & Neuberg, 1990). It also aligns with arguments in literary essays that fiction can foster attention to the specificity of persons and situations, making it harder to reduce individuals to generic types (Brodsky, 1987; Keen, 2007; Nussbaum, 1990; Zunshine, 2006). Importantly, individuation is not the same as self–other overlap (Aron et al., 1992): instead of making others feel closer to the self, it predicts greater attention to the specific features that differentiate individuals and weaker reliance on categorical heuristics—ultimately, the kind of attention Brodsky (1987) evokes when an empty circle (a zero) is transformed, by literature, into a human face: not necessarily pretty, nor even fully comprehensible. Motivated by these ideas, our measurement strategy combines standard measures of empathy, immersion and closeness (Transportation and the Inclusion of Other in the Self scale, IOS) with a perceived uniqueness/individuation scale.

A simple way to understand this mechanism is to think of prosocial action as depending on the weight an individual places on another person's payoff. In that view, fiction need not increase prosociality by uniformly making others feel closer or more similar. Instead, due to the richness of stories and psychologically textured characters, it may change the distribution of social weights by making some image of others more behaviorally salient in ways they otherwise would not be. Put differently, literary fiction may reduce the tendency to treat others as generic types and increase the probability that they are perceived as psychologically specific individuals whose welfare enters decision-making. This framework implies that prosociality can rise together with greater attention to individual differences, more complex representations of others' behavioral motives, and less reliance on simple interpretations of others, even if self–other overlap does not increase.

Our main empirical finding is that exposure to literary fiction increases prosocial behavior in economically meaningful ways. Relative to the placebo course, participants assigned to the fiction-reading curriculum give more in the dictator game and return a larger share as trustees in the trust game. By contrast, we do not detect an increase in trusting behavior by first movers, and no robust effects in the ultimatum or third-party punishment game, where strategic considerations and counter-punishment are salient. Taken together, the behavioral pattern is one of greater altruism and greater trustworthiness, rather than a generalized change in strategic behavior. In line with the conceptual framework above, this is consistent with fiction changing the conditions under which another person's welfare enters the choice problem.

A second result concerns the source of these behavioral changes. Using incentive-compatible elicitation of personal norms, empirical and normative social expectations, we find that the treatment effects are more consistent with shifts in individual preferences than with broad updating of perceived social norms. In particular, fiction-treated students become more generous and more trustworthy even though we do not observe changes in second-order beliefs about how others behave. At the same time, they become somewhat more permissive in their own judgment about the lack of trust and expect others to be more permissive regarding unfair behavior in the third-party punishment game. This suggests

that the intervention does not operate through tightening of prosocial norms; if anything, it points to a mechanism in which participants place greater behavioral weight on others welfare while also adopting a less disapproving attitude toward deviations from what is considered normative behavior.

Our mechanism results are consistent with narrative immersion fostering social imagination and, in turn, a greater valuation of others' individuality, rather than with an increase in moral self-perception, empathy, or closeness to others. First, we find evidence that the fiction-reading course leads to higher narrative immersion and transportation into stories, which is consistent with the course developing participants' ability to read better, engage more cognitively with social stories, and imagine more fully the world being narrated. However, we find no change in either affective or cognitive empathy. Second, the IOS and social connectedness patterns do not support a "closeness" account: if anything, treated participants report lower overlap with several others, especially when those others are socially unfamiliar. When we examine more closely what might explain this increase in perceived distance, we find treated participants are more likely to form multifaceted explanations of human behavior. In particular, they are more aware of the interdependencies and causes underlying others' behavior, more comfortable with the ambiguity of others' motivations, and more likely to recognize that fully understanding another person requires effort. In short, other's behavior becomes less reducible to default interpretations. Finally, a new scale developed in the paper to measure the extent to which participants value the uniqueness of others provides suggestive evidence that fiction increases attention to what differentiates one person from another. These findings support the idea that literary fiction may foster prosociality not by making others feel more like the self, but by making them harder to reduce to a generic type and easier to see as psychologically specific individuals. Linking this back to the theoretical framework, fiction-reading makes the social representation model of others richer, and their welfare becomes more relevant in determining one's own utility.

Our paper contributes to several strands of the literature. The first contribution is methodological, we complement standard self-report outcomes with a revealed-preference approach. Much existing work relies on questionnaires administered during or immediately after reading, which makes it difficult to disentangle changes in preferences and norms from priming or social desirability. We therefore measure prosocial behavior using incentivized economic games, and we separately elicit social norms using the coordination-based method of Krupka and Weber (2013). This framework allows us to evaluate not only whether prosociality changes, but also whether any change is better interpreted as a shift in personal preferences or as an updating of empirical and normative social expectations. Using this approach, we find that participants in the fiction-reading treatment sent more in the dictator game and returned a larger share as trustees in the trust game than those in the control group, indicating increased altruism and trustworthiness (with no significant change in trusting behavior as first movers). They also relaxed their personal norm against returning less than half in the trust game and believed that others would view such low returns as more acceptable.

Second, our design helps speak to persistence. Participants engage with the intervention

across multiple sessions and complete behavioral assessments beyond the immediate reading context, which reduces the likelihood that results are driven purely by short-lived priming effects. This matters because the persistence of fiction-reading effects remains an open question in the broader literature (Dodell-Fedder & Tamir, 2018). In our data, the behavioral effects remain significant even at the four-week follow-up, whereas some self-reported effects—particularly the increase in attention to individual differences—are larger at the immediate measure, suggesting that certain perceptual shifts may emerge immediately rather than consolidate gradually. This temporal pattern reinforces the value of designs that extend measurement beyond the intervention window.

Additionally, recent work in economics has emphasized that prosocial behavior is not a fixed trait but a malleable component of preferences shaped by information and experience. A leading example is Andries et al. (2024), who study empathy formation using immersive virtual reality. They show that combining neutral information with an emotionally salient experience increases charitable giving and pro-immigration attitudes, and argue that this effect operates through increased perceived similarity: individuals become more empathetic when they are able to experience others' hardship as if they were "in their shoes." One should note that it can work in the opposite direction, increasing xenophobia (Braun, 2022). In a complementary framework, Enke (2025) conceptualizes prosociality in terms of moral universalism— understood as to what degree a person's altruism is invariant to the identity of others— documenting systematic heterogeneity in how broadly individuals extend moral concern beyond in-groups and find it is correlated with historical economic incentives such as tightness of kinship networks and irrigation practices.

A related strand of the literature highlights the role of beliefs and narratives in shaping prosocial attitudes and policy preferences. Alesina et al. (2018) and Stantcheva (2022) show that support for redistribution and generosity depends critically on beliefs about effort, luck, and deservingness, and that narratives and information can durably affect these beliefs. Alan et al. (2023) show that strengthening meritocratic beliefs increases effort while reducing altruism toward individuals perceived as responsible for their outcomes. Historical and educational evidence further suggests that exposure to morally charged narratives can induce persistent value change (Figueroa & Fouka, 2023; Hartinger et al., 2026; Natalia Amaya & Hosie, 2024). Our contribution differs from these approaches by identifying a distinct cognitive mechanism. Rather than increasing empathy, perceived similarity, or the scope of moral concern, exposure to literary fiction increases attention to individual differences, weakening categorical representations of others. This individuation channel operates independently of responsibility attribution and belief updating, and provides a micro-foundation for how narrative experiences can shape prosocial preferences. The work of Nagel (1970) develops a similar hypothesis, arguing that "altruism itself depends on a recognition of the reality of other persons and on the equivalent capacity to regard oneself as morally one individual among many."

Finally, this paper relates to broader work on educational interventions that seek to shift socially beliefs and behaviors at early ages (Alexander Cappelen & Tungodden, 2020). One approach in this domain is to design curricula that explicitly develop perspective-taking and

prosocial values. Our intervention aligns with this agenda by treating fiction as a laboratory for developing socio-emotional skills relevant to prosociality, and by pairing the curriculum with incentivized measures of prosocial behavior. In doing so, we bridge insights from the humanities and behavioral science, illustrating that literary education can measurably influence how people act toward others.

2 Conceptual framework

A simple way to organize our interpretation is to model prosocial behavior as depending on a latent social weight placed on the other person's payoff which is informed by the social representation model they have about others.[1] Suppose an individual chooses a prosocial action a that is costly to the self and beneficial to another person. Let utility be

$$u(a \mid \beta) = -c(a) + [\beta]_+ b(a), \qquad [\beta]_+ \equiv \max\{\beta, 0\},$$

where $c(a)$ is the cost of prosocial action, $b(a)$ is the benefit to the other, and β captures the extent to which the decision-maker takes the other's welfare into account, which is drawn from a social representation model s given by distribution F_s. When $\beta \leq 0$, the other's payoff does not enter utility and behavior coincides with the selfish benchmark. When $\beta > 0$, the other's payoff becomes behaviorally relevant and affects optimal choices.

We interpret fiction not necessarily as increasing the average level of concern for others, but as changing the distribution of this social weight. In particular, fiction may make social representations richer, so that individuals who would otherwise be treated as generic or interchangeable are instead processed as distinct. In the model, this corresponds to fiction-reading induced social representation model F_T becomes a mean-preserving spread of the social model in the absence of fiction-reading F_C: treatment increases dispersion in how much weight is placed on others without necessarily shifting the mean.

This framework yields testable implications for behavior and mechanisms. First, under standard curvature assumptions (convexity of the cost and concavity of the other person's payoff), a richer social representation model increases expected prosociality, because the optimal action $a^*(\beta)$ is convex in β, even in the absence of changes in social norms (see Proposition (1)). Second, if richer social representation reduces reliance on generic interpretations of others, this can be understood in the model as expanding the mapping from β to the set of reasons attributed to others' behavior. Accordingly, the fiction treatment should be associated with finer categories people use to interpret others, greater attributional complexity, and more dispersion in perceived motivations (Proposition (2)). Third, the framework does not require higher self–other overlap: if fiction works by making others more distinct rather than more similar, then prosociality may increase even when personal feelings of closeness to others remain unchanged or decline and scales related to attention to individuality increase (Proposition (3)).

[1] The formal version of this argument is found in Appendix A.2.

3 Experimental design

All the design was preregistered in the AEA RCT Regisry (AEARCTR-0015996). Participants were recruited via an email invitation to participate in a course about behavioral science, literature and its applications to daily life [2]. This invitation was purposely wide to include both our treatment and placebo courses in a way that no participant could feel deceived, reducing both attrition and selection.

The experiment was conducted in two waves with a total of 344 student participants ($N = 344$) recruited at Universidad de los Andes (Bogotá, Colombia). The first wave ($n = 58$) ran from late April to May 2025 and drew participants from the Economics Department's online laboratory pool while the second wave ($n = 286$) ran from October to November 2025 and was open to undergraduates across the university. Both waves included the same four one and a half hour sessions of guided reading. Before the first session all participants were asked to fill an Entrance Survey with demographic and question about their reading habits and opinion on the effect of fiction reading in attitudes. Participants were randomly assigned (in roughly equal proportions) to one of two study arms (literary-fiction treatment vs. behavioral-science placebo).

The treatment arm consisted of a literary fiction reading course titled *"Mejora Tu Lectura Literaria"*, developed in collaboration with Paredro Leo and facilitated by a Ph.D. in Literature. This course is mainly an introductory course focused on cultivating fiction reading as a habit. Participants read and discussed selections of Colombian fiction (novel excerpts and short stories in Spanish), with an emphasis on understanding each character as a unique individual and reflecting on characters' emotions and perspectives. As its names suggests, the course was intended to make participants better readers that could move from just a mere evaluative reading in which the reader can only say if it enjoys the story told to a more involved one, in which there is a deep personal identification with the reading material. Of course, text where designed to generate this transition with fiction materials that where close to the lived experience of participants[3]. As we will develop later on, we are not arguing that any curriculum of fiction readings can increases prosociality. We are arguing that a well chosen set of reading materials read with the guidance of an expert can have tangible effects on prosociality.

The first session is guided by two questions: Why we read? and What do we read for? (In Spanish:*¿Por qué y para qué leemos?*). A classical text by Louise Mary Rosenblatt[4] is

[2]The full text sent in Spanish is as follows: *La Facultad de Economía de la Universidad de los Andes te invita a participar en un curso sobre ciencias del comportamiento, literatura y aplicaciones en la vida diaria.* Figure A.1 shows the actual email sent

[3]As one person commented during one of our presentations: the selected works are *hard to read*, in the sense that they involve talking about the history of political violence that Colombia has had in recent decades

[4]The text assigned is the Spanish translation, naturally:*La literatura como exploración*, Louise Mary Rosenblatt, FCE, 2003. The original text is Literature as Exploration (New York: D. Appleton Century Co., 1938)

used as a canonical conception of what literature *can* do. The firs of the seven possibilities discussed is the creation of the necessary moral imagination that democracy need in its citizens Figure A.18, and that is presented as the basis for the rest of the course. A few examples of texts are discussed in order to discuss what elements make a text literary. The second session of the treatment arm is based on the premise that the explicit goal of literature is to allows us to know the other from the inside. In this session the discussion is centered on what makes for a good reader, and thus for good literature. This is translated to the discussion of the elements of a good literary piece, and the course start by discussing the different types of literary narrators, and the initial paragraphs of some classical novels (*Lolita*, *Infinite Jest*) are used as examples. The third session is dedicated to further enriching the notion of increasing empathy through literary reading. The role of imagination is emphasized as the vehicle that permits the reader to connect with others even without having the same experiences themselves. In this session literature is put in dialogue with other art forms, particularly photography and painting (examples used are related to the representation of political violence also seen in the literary texts and the, see Figure A.19). Remember, that in the third session a quiz of reading comprehension and a prize is given for the higher scores. The fourth session is a summary session that connects all the text discussed with recent advances in neurosciences and the explicit idea that fiction reading can create mindfulness. This is shown in connection to the work of Richard J. Davidson, and the course shows how reading can create Awareness, Insight, Connection and Purpose.

In contrast, the control arm received a placebo course titled *"Curso de Ciencias del Comportamiento"*, taught by two Ph.D. economists. The placebo curriculum covered fundamentals of behavioral science—topics such as decision theory, cognitive biases, social preferences, nudges, and behavioral interventions. Readings for the control course were drawn from general-audience texts in behavioral economics and psychology (e.g., popular excerpts from Kahneman and Thaler), with no fictional narratives or literary immersion included. This course was also divided in four sessions.

The first session gives a brief history of behavioral economics as an independent area of research. Some of the classical work of Kahneman and Tversky on risk aversion is discussed as well as the work of Read and van Leeuwen (1998) on self control. The second session of the course is more focused on the practical application of these findings. A significant amount of the session is dedicated to research on the planing fallacy, and proven strategies to reduce its effects in daily life. The rest of the session introduces the main elements of prospect theory as a rebuttal of classical Expected Utility Theory. Third session introduces the notion of non-individual and non-monetary incentives and discusses profusely recent research that studies this question, particularly the work of Gneezy and Rustichini (2000). Some practical applications of his research agenda to companies are also discussed. Fourth and final session focuses on recent research on Social Norms. A brief history of this area of study is shown before discussing applications, like the work of Leonardo Bursztyn and Yanagizawa-Drott (2020) around social norms Saudi Arabia. The course was designed not to cover any of the experimental games in detail, nor was any discussion of the existence of them in the final session.

A guide to the content of each course can be seen in the Appendix of this paper, and the full course materials are available on the accompanying texts. Sessions were held in small-group seminar format, with an average of 45 students per session. In the third session, participants completed an incentivized quiz on the course reading material (to encourage thorough engagement). Key outcome measures were collected at the end of the final session (immediately post-intervention) and then again in a follow-up session approximately four weeks later. Note that both the treatment and the placebo arm are bundled treatments in which there are guided discussions of the material with an expert on the field. It is guided reading, not reading in itself that can have effects on behavior.

3.1 Incentivized experimental games and norm elicitation

Immediately after the final course session (short-term session), or in a follow-up session four or eight weeks later (long-term session), participants completed a set of incentivized economic games that capture complementary components of prosociality—altruism, trust, trustworthiness, strategic fairness, and norm enforcement. Subjects faced half of the games in the session immediately after the course and the rest in the follow-up (we randomly vary this order). Decisions were anonymous and implemented for money using a token system. Participants in both treatment conditions were compensated if they participated in at least three course sessions (USD22 approx.), and they could earn additionally a maximum of USD16.5 for their decisions in each of the experimental sessions.

Dictator game. To measure altruism, we implemented a standard dictator game (Forsythe et al., 1994). In the role of *Player 1*, participants decided how many tokens (0–10) to transfer to an anonymous *Player 2*, keeping the remainder. In the role of *Player 2*, participants stated their incentivized *empirical expectation* about the most common transfer made by Player 1 participants.

Trust game. To measure trust, trustworthiness, and reciprocity, we implemented the canonical investment (trust) game (Berg et al., 1995). As *Player 1* (trustor), participants chose how many tokens (0–10) to send to *Player 2*. The amount sent was multiplied threefold before reaching Player 2. As *Player 2* (trustee), participants made return decisions using the strategy method for all possible amounts sent by Player 1. We also elicited incentivized Player 1 beliefs about how much other trustors typically send, and Player 2 beliefs about the most common amount sent by trustors.

Ultimatum game. To measure strategic fairness and acceptance thresholds, we implemented the ultimatum game (Güth et al., 1982). As *proposer* (Player 1), participants chose an offer (0–10) from 10-tokens. As *responder* (Player 2), participants indicated whether they would accept or reject each possible offer, using the strategy method, which allows

us to recover a typical minimum acceptable offer. We elicited incentivized beliefs about the most common offer made by proposers and about the typical responder acceptance threshold.

Third-party punishment game. To measure fairness norm enforcement, we implemented a third-party punishment paradigm (Fehr & Fischbacher, 2004). Participants made decisions both as an allocator (Player 1) in a dictator-type allocation to Player 2 and as an observer (Player 3) who could spend tokens to reduce Player 1's payoff after observing the allocation, under a fixed punishment technology where one token allocated to punishing Player 1 reduced their payoffs by three.Using the strategy method, Player 3 specified punishment levels contingent on all possible allocations. We also elicited incentivized beliefs about typical punishment as well as the most common amount sent by Player 1.

Personal norms and normative expectations. To distinguish changes in private preferences from changes in perceived social norms, we elicited both self-reported *personal norms* and *normative social norms*. For the former, participants rated how appropriate they considered a focal deviation (e.g., giving less than half in the dictator game; sending less than half in the trust game; punishing "too little" in the third-party game) on a 1–4 scale from "very inappropriate" to "very appropriate". For eliciting normative social norms, we followed the coordination-based method of Krupka and Weber (2013). For each relevant behavior, participants rated how *socially* appropriate they considered the same deviation and were told they would be paid if their response matched what other random chosen participant would answer for the same question.

3.2 Psychological scales

To shed light on mechanisms, participants also completed psychological scales designed to capture immersion and social-cognitive mechanisms emphasized in the fiction-reading literature (Fiske & Neuberg, 1990; Green & Brock, 2000a). In specific, we used a Transportation Scale based on , the Narrative Engagement Scale, the Inclusion of Other in the Self (IOS), Moral Identity Scale, the Attributional Complexity Scale, the Moral Identity Scale and a revised Reading the Mind in the Eyes Test were administered.

We also designed a Perceived Uniqueness scale developed for this study, informed by the individuation-versus-categorization framework, to test some mechanisms that we wanted to test in line with our conceptual framework. All psychological scales will be discussed in detail when presenting results. In order to be concise, we focus on significant results.

4 Main results

In this section we present the main empirical results of the experiment. We begin by noting that observable characteristics are statistically similar between the Treatment and Placebo groups Table 1. As expected given the recruitment setting, the sample is relatively young and disproportionately drawn from upper socioeconomic strata, which is natural for a private university in Bogotá. The entrance survey completed before the intervention included participants' pre-treatment reading habits to account for pre-existing differences. The treatment group reports slightly higher average reading hours, but this does not translate into a higher number of books read or purchased during the previous year.

Table 1: Sample balance. Demographics.

	(1) Mean	(2) Sd	(3) Treated	(4) Placebo	(5) pval
Age	19.97	2.29	20.18	19.74	0.08
First wave (1=first wave, 0=otherwise)	0.17	0.38	0.15	0.19	0.34
Short term (1=immediate participation, 0=otherwise)	0.62	0.49	0.63	0.61	0.71
Knowother (1=knows another participant, 0=otherwise)	0.20	0.40	0.23	0.17	0.19
Female (1=female, 0=otherwise)	0.58	0.49	0.59	0.57	0.65
High strata (1=5 and 6, 0 =1,2,3,4 strata)	0.15	0.36	0.14	0.17	0.49
Econ field (1=economics related field, 0=otherwise)	0.17	0.38	0.17	0.17	0.98
Quiz score (correct answers; min=0, max=86)	19.59	13.58	17.88	21.49	0.01
Fiction read (# of fiction books read)	2.81	4.11	3.12	2.48	0.16
Observations			183	161	

Notes: This table reports baseline balance between participants assigned to the fiction-reading treatment and the placebo condition. Columns (1)–(2) show the overall mean and standard deviation in the full sample. The full sample is defined as individuals who attended at least three sessions and at least one experimental session. We additionally report results conditional on participation in at least one experimental session, where participation is measured by the availability of non-missing information for at least one game outcome or one scale measure. Columns (3)–(4) report group means by assignment status. Column (5) reports two-sided p-values from tests of equality in means between the treated and placebo groups for each covariate. Indicator variables include whether the participant took part in the first wave, whether the participant attended the immediate post-course session (short-term), whether the participant was acquainted with someone from the opposite course, whether the participant is female, whether the participant's household utility bill corresponds to socioeconomic strata 5 or 6 (high strata), and whether the participant studies economics, business, or industrial engineering (econ field). Age is measured in years. Quiz score is the pre-treatment performance measure from the Author Recognition Test. Fiction read denotes the number of fiction books read.

A particularly useful measure is the Author Recognition Test (ART), which we adapted to the local context and made incentive-compatible by paying a USD16 monetary prize to the two-highest scoring participants. The test consists of identifying literary authors from a list that also includes names of well-known non-authors; participants receive one point for each correct answer and lose one point for each incorrect answer. To construct the list, we used the

most searched authors in Bogotá's public libraries.[5] As the table shows, the placebo group performs slightly better on this measure, if anything working against our prior hypothesis that treatment participants may have entered the study with greater literary familiarity.

We also measure participants' prior beliefs about the effects of fiction reading. Respondents evaluated a series of statements about whether reading fiction changes how people think, feel, or behave. A useful feature of these data is that they allow us to assess whether our results might simply reflect the reinforcement of pre-existing beliefs about the virtues of fiction. Overall, participants show only mild agreement with the idea that fiction has positive effects on people's minds (see Table A.10). In that sense, it is unlikely that our findings are driven purely by the confirmation of strongly held prior views. It is particularly notable that, contrary to what literary theorists often emphasize, participants are not especially convinced that reading fiction makes people more empathetic. If anything, they seem somewhat more inclined to associate fiction with greater psychological complexity than with empathy per se.

4.1 Prosocial behavior and social norms

We now turn to the treatment effects on behavior and social norms. Throughout this subsection, outcomes are regressed on an indicator for assignment to the fiction-reading curriculum, relative to the placebo condition. For precision, all specifications include controls for wave, the session in which the measure was collected, and the baseline characteristics reported in Table 1.

The results reveal that exposure to literary fiction increases prosocial behavior in the dictator game and in the trustee role of the trust game, but we do not find comparable effects on trusting behavior by first movers, nor robust effects in the ultimatum or third-party punishment games. Taken together, this pattern is more consistent with an increase in altruism and trustworthiness than with a generalized shift in strategic behavior.

To estimate treatment effects on the main behavioral outcomes, we use the following specification:

$$y_{i,t} = \alpha + \beta \, \text{Fiction}_i + \pi_{w(i)} + \mathbf{x}_i'\boldsymbol{\gamma} + \varepsilon_{i,t}, \tag{1}$$

where $Y_{i,t}$ denotes the behavioral outcome of interest for individual i in experimental session-term t (either immediately after the course finished or in a four-week or eight-week follow-up), such as the amount sent in the dictator game, the amount sent by the first mover in the trust game, the ultimatum offer, or the amount deducted in the third-party punishment game. α is a constant term. Fiction_i is an indicator equal to one if participant i was assigned to the fiction-reading curriculum and zero if assigned to the placebo condition. The coefficient of interest, β, therefore captures the average treatment effect of fiction exposure on the corresponding behavioral outcome. $\pi_{w(i)}$ denotes wave fixed effects. $\mathbf{x}_i$ is a vector of baseline covariates, including the characteristics reported in Table 1, and $\boldsymbol{\gamma}$ is the associated vector of

[5] As shown in Figure A.3, both groups of names (writers and non-writers) are balanced in terms of gender and foreign origin, and they contain the same number of deceased individuals.

coefficients. Finally, $\varepsilon_{i,t}$ is an idiosyncratic error term. In all specifications, standard errors are clustered at the treatment-session-term level.

For outcomes observed using the strategy method, we stack the decisions made by each participant across the relevant within-game actions of Player 1 and estimate:

$$y_{i,j,t} = \alpha + \beta\,\text{Fiction}_i + \sum_{k\neq 0} \delta_k \mathbf{1}\{j=k\} + \sum_{k\neq 0} \theta_k\,(\text{Fiction}_i \times \mathbf{1}\{j=k\}) + \pi_{w(i)} + \mathbf{x}_i'\boldsymbol{\gamma} + \varepsilon_{i,j,t}. \quad (2)$$

$y_{i,j,t}$ denotes the behavioral outcome of interest for individual i, at decision point or scenario j, and experimental session-term t. This specification is used for outcomes that vary within participant across possible game contingencies. In the trust game, $y_{i,j,t}$ corresponds to the percentage of the received amount sent back by the trustee when Player 1 sends j points; in the ultimatum game, it corresponds to the acceptance decision for offer j; and in the third-party punishment game, it corresponds to the punishment assigned for allocation or amount-received condition j.

The index j refers to the strategy-method potential behavior of Player 1, while k indexes the set of possible actions included in the summation. Thus, $\mathbf{1}\{j=k\}$ is an indicator equal to one when the observed decision corresponds to Player 1 sent amount k. The omitted contingency is the reference category when Player 1 sends 0 tokens. The coefficients δ_k capture how the outcome differs across potential Player 1 decisions in the placebo group, relative to the reference category. The coefficient β captures the treatment effect of fiction when Player 1 sends 0 tokens, while the interaction coefficients θ_k indicate whether the treatment effect differs when Player 1 sends k tokens. Hence, for any $k \neq 0$, the estimated treatment effect of fiction is $\beta + \theta_k$. This specification therefore allows fiction to affect both the level of the strategy-method response and the shape of the response profile across Player 1 decisions.

As before, α is a constant term, Fiction_i is an indicator equal to one if participant i was assigned to the fiction-reading curriculum and zero if assigned to the placebo condition, and $\pi_{w(i)}$ denotes wave fixed effects. The vector $\mathbf{x}_i$ contains baseline covariates, including the characteristics reported in Table 1, and $\boldsymbol{\gamma}$ is the associated vector of coefficients. Finally, $\varepsilon_{i,j,t}$ is an idiosyncratic error term. In all specifications, standard errors are clustered at the treatment-session-term level.

4.1.1 Dictator Game

Tables 2 and 3 present the results for the dictator game, our cleanest measure of altruistic behavior. The treatment effect on Player 1 transfers is positive and economically meaningful: participants assigned to the fiction-reading course send around 12% more of their initial endowment. This increase is not mirrored by comparable changes in empirical expectations or in elicited social norms, suggesting that the behavioral effect is not primarily driven by updated beliefs about what others do or by a tightening of prosocial norms.

If anything, this pattern is more consistent with a direct shift in individual preferences. Moreover, when we include additional observations for which some demographic controls are missing, both the magnitude and significance of the treatment coefficient become somewhat stronger, reinforcing the interpretation that the fiction curriculum increased altruistic giving.

Table 2: Player 1 (Altruistic) Behavior and Empirical Expectations

	Amount sent		Expectations	
	(1)	(2)	(3)	(4)
Treatment	0.678**	0.501**	0.335	0.260
	(0.271)	(0.227)	(0.282)	(0.192)
Constant	3.615***	3.619**	3.498***	2.361
	(0.133)	(1.391)	(0.250)	(1.426)
Placebo mean	3.614	3.614	3.506	3.506
Observations	175	171	175	171
R^2	0.033	0.121	0.018	0.133
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of dictator-game outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are (i) the number of tokens sent by Player 1 (*Amount sent*) and (ii) Player 2's incentivized empirical belief about the most common transfer chosen by Player 1 (*Belief about amount sent*). The dictator game was implemented with a 10-token endowment. In the belief-elicitation task, participants received a monetary bonus if their guess coincided with the modal transfer among Player 1 participants. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.1.2 Trust game

Table 4 present the results for the trust game when participants act as Player 1. In contrast to the dictator game, we do not detect a robust treatment effect on the amount sent. In other words, the fiction-reading intervention does not appear to increase trusting behavior in a context where choices depend both on prosocial inclinations and on strategic beliefs about the partner's likely response.

At the same time, we report in Table A.11 the personal norms and normative expectations from Player 2 regarding non-trusting behavior of Player 1 the norm measures reveal an interesting shift. The coefficients are coded so that higher values indicate greater personal (columns 1 an 2) and social (columns 3 and 4) permissiveness toward low-trust behavior,

Table 3: Player 2 Personal and social norms regarding selfishness

	Personal norm		Normative expectations	
	(1)	(2)	(3)	(4)
Treatment	-0.047	-0.024	-0.014	-0.022
	(0.101)	(0.096)	(0.117)	(0.112)
Constant	2.659***	2.831***	2.190***	2.071***
	(0.075)	(0.389)	(0.097)	(0.552)
Placebo mean	2.663	2.663	2.193	2.193
Observations	175	171	175	171
R^2	0.016	0.050	0.008	0.062
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of norm-related outcomes in the dictator game on an indicator for assignment to the fiction-reading treatment. The dependent variables are (i) participants' personal norm regarding sending less than half of the endowment in the dictator game, and (ii) perceived social norms regarding the same behavior. Personal norms are measured on a 1–4 scale ranging from "very inappropriate" to "very appropriate," capturing respondents' own evaluation of the action. Social norms are elicited using the coordination-based method of Krupka and Weber (2013), in which participants receive a monetary bonus if their response matches the modal answer among other participants. The dictator game was implemented with a 10-token endowment. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include the characteristics reported in Table 1. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

and treated participants appear somewhat more willing to judge such behavior as acceptable. This is important because it suggests that the intervention did not operate through a general tightening of prosocial norms. If anything, fiction-treated participants become more prosocial in some domains while also becoming less disapproving of certain deviations from normative behavior.

On the other hand fiction-treated Player 2 return a larger share of the amount received, indicating greater trustworthiness. Figure 1 shows that this treatment effect reflects a shift in the entire return schedule estimated from the strategy-method data: relative to the placebo group, treated participants display a higher intercept, indicating greater baseline trustworthiness, but a flatter slope, indicating less reciprocity in the narrow sense of responding more strongly to higher levels of trust. Both features are statistically significant. By contrast, there is no comparable effect on personal judgments or on beliefs about the social appropriateness of returning less, suggesting that the behavioral change is not mainly operating through shifts in perceived norms. Overall, the intervention appears to increase

Table 4: Player 1 (Trust) Behavior and Empirical Expectations

	Amount sent		Expectations	
	(1)	(2)	(3)	(4)
Treatment	-0.328	-0.412	-0.066	-0.322
	(0.426)	(0.401)	(0.270)	(0.775)
Constant	6.458***	9.694***	5.561***	13.116**
	(0.291)	(2.639)	(0.168)	(4.547)
Placebo mean	6.470	6.470	5.566	5.566
Observations	175	171	175	171
R^2	0.014	0.057	0.003	0.046
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of trust-game outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are (i) the number of tokens sent by Player 1 (trustor) to Player 2, and (ii) Player 1's incentivized empirical belief about the typical amount sent by other trustors. In the trust game, Player 1 decided how many tokens (0–10) to send to Player 2. The amount sent was tripled before reaching Player 2. Beliefs were elicited using an incentivized coordination scheme: participants received a monetary bonus if their response matched the modal answer among other participants. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

unconditional trustworthiness more than conditional reciprocity.[6]

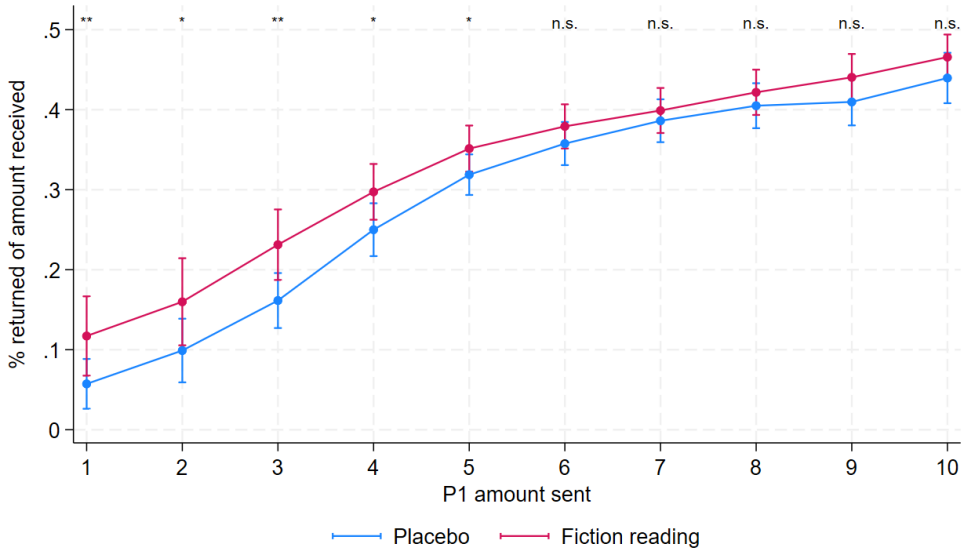
[6]See Table 6 for the corresponding regression results and Table A.11 for personal and normative norms from Palyer 1 regarding trustworthiness.

Table 5: Player 2 Personal and Social Norms Regarding Non-Trusting Behavior of P1

	Personal norm		Normative expectations	
	(1)	(2)	(3)	(4)
Treatment	0.252**	0.243***	0.214**	0.174**
	(0.095)	(0.077)	(0.085)	(0.059)
Constant	2.108***	2.887***	1.853***	2.130***
	(0.057)	(0.651)	(0.022)	(0.468)
Placebo mean	2.108	2.108	1.855	1.855
Observations	175	171	175	171
R^2	0.021	0.064	0.025	0.066
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of trust-game norm-related outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are (i) participants' personal norm regarding sending fewer than five tokens in the trust game, and (ii) perceived social norms regarding the same behavior. Personal norms are measured on a 1–4 scale ranging from "very inappropriate" to "very appropriate," capturing respondents' own evaluation of sending fewer than five tokens. Social norms are elicited using the coordination-based method of Krupka and Weber (2013), in which participants receive a monetary bonus if their response matches the modal answer among other participants. In the trust game, Player 1 decided how many tokens (0–10) to send to Player 2. The amount sent was tripled before reaching Player 2. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure 1: Percentage returned of amount sent in the Trust Game by treatment status.



Notes: The figure plots the mean percentage of tokens returned by Player 2 (trustee) in the trust game, separately by treatment status and by the amount sent by Player 1. The dependent variable is the share of the tripled amount returned to Player 1, constructed as $Y/(3X)$, where X denotes the number of tokens sent by Player 1 and Y the number of tokens returned by Player 2 under the strategy method. Points correspond to fitted values from a fully saturated model interacting treatment assignment with the amount sent by Player 1 and including fixed effects for the first experimental wave. Vertical bars denote 95% confidence intervals. Standard errors clustered at the treatment session term level are reported in the confidence intervals. Significance markers above each amount indicate tests of equality between the fiction-reading treatment and placebo at that specific amount sent: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, and n.s. indicates not statistically significant at the 10% level.

Table 6: Player 2 (Trustworthiness) Behavior and Empirical Expectations

	% sent back (P2)		expected % sent back	
	(1)	(2)	(3)	(4)
	Baseline	With controls	Baseline	With controls
Treatment	0.068**	0.057*	0.100**	0.094*
	(0.034)	(0.033)	(0.045)	(0.045)
Amount sent (P1)	0.044***	0.044***	-0.010**	-0.010**
	(0.003)	(0.003)	(0.004)	(0.004)
Treatment × Amount sent (P1)	-0.005	-0.005	-0.012*	-0.011
	(0.004)	(0.004)	(0.007)	(0.006)
Constant	0.046**	0.032	0.355***	0.381***
	(0.020)	(0.024)	(0.032)	(0.035)
Placebo mean	0.288	0.288	0.291	0.292
Observations	1380	1350	172	168
R^2	0.398	0.422	0.140	0.165
Wave fixed effects	✓	✓	✓	✓
Demographic controls		✓		✓

Notes: Entries report coefficients from linear regressions of trust-game outcomes on an indicator for assignment to the fiction-reading treatment, the amount sent by Player 1, and their interaction. The dependent variable in Columns (1)–(2) is the percentage returned by Player 2, defined as the share of the tripled amount received from Player 1 that Player 2 sends back. The dependent variable in Columns (3)–(4) is Player 1's expectation about how much Player 2 will return. In all columns, *Amount sent (P1)* refers to the observed amount sent by Player 1. In the trust game, Player 1 decided how many tokens (0–10) to send to Player 2, and the amount sent was tripled before reaching Player 2. Return behavior and return expectations are therefore interpreted conditional on the amount initially sent by Player 1. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.1.3 Ultimatum and Third-party punishment games

The ultimatum game results provide little evidence of a robust treatment effect.[7] Neither proposer behavior nor responder acceptance patterns show systematic changes across treatment conditions. This null result is informative. Unlike the dictator game, ultimatum behavior reflects not only distributive preferences but also strategic anticipation and bargaining considerations. The absence of clear effects therefore reinforces the broader pattern of the paper: fiction exposure appears to affect altruism and trustworthiness more than behavior in environments where strategic considerations are central.

The third-party punishment game also yields no robust treatment effects on fairness by Player 1 or on punishment of unfairness from Player 3.[8] Nor do we observe systematic changes in empirical expectations regarding these behaviors or normative expectations regarding unfair behavior. If anything, the most notable effect is that treated participants appear somewhat more permissive in their own evaluations of not sanctioning unfair behavior.[9]

This matters for interpretation. If the fiction-reading intervention had operated by strengthening prosocial norm enforcement, we would expect harsher punishment of selfish behavior. That is not what we find. Instead, the results suggest a more nuanced mechanism: participants in the fiction arm place greater behavioral weight on others' welfare in some settings, while becoming less punitive or less morally rigid in their judgments about deviations from normatively expected behavior.

4.2 Mechanisms

We next examine the psychological scales to better understand the mechanisms behind the behavioral effects. The results are most consistent with narrative immersion and individuation, rather than with increases in empathy, moral self-perception, or closeness to others. Fiction may increase prosociality not by making others feel more similar to the self, but by making them more psychologically specific and behaviorally salient. The evidence below largely supports that interpretation.

For scale-based outcomes, we estimate:

$$S_i = \alpha + \beta \, \text{Treatment}_i + \pi_{w(i)} + \mathbf{X}_i' \Gamma + u_i, \tag{3}$$

where S_i denotes the score of individual i on the corresponding scale or index, such as attributional complexity, empathy, uniqueness, loneliness, or moral identity. α is a constant

[7] See Table A.12 for Player 1 offers and Figure A.6 and Table A.14 for Player 2 acceptance behavior in Appendix

[8] See Table A.17 and Figure A.7 in Appendix.

[9] See Table A.17 and Table A.18.

term. Treatment_i is an indicator equal to one if participant i was assigned to the fiction-reading curriculum and zero otherwise, so that β captures the average treatment effect on the relevant scale outcome. $\pi_{w(i)}$ denotes wave fixed effects, implemented through an indicator for the first experimental wave. $\mathbf{X}_i$ is a vector of baseline covariates where we included age, ATR quiz score, firstwave, short term, if the person knew others, female, high stratum and if they where studying an economics related field, Γ is the associated vector of coefficients, and u_i is an idiosyncratic error term. As in the behavioral specifications, standard errors are clustered at the treatment-session-term level.

4.2.1 Narrative engagement and immersion

Narrative transportation. To capture broader engagement with narratives beyond transportation—including attentional focus, emotional involvement, and perceived realism—we administered a Narrative Engagement scale (Busselle & Bilandzic, 2009). The instrument draws on the Transportation framework of Green and Brock (2000b) and the multidimensional Narrative Engagement model of Busselle and Bilandzic (2009). Items are answered on a Likert scale and aggregated (after reverse-coding where applicable) so that higher values indicate greater engagement with the narrative experience. Participants completed the scale after reading the short story *"El hijo"* by Horacio Quiroga, allowing us to test whether the fiction-reading curriculum increased the extent to which participants felt absorbed by a narrative (see Table A.5).

Participants assigned to the fiction-reading course report higher levels of transportation, cognitive accessibility, narrative presence, involvement, and loss of self and time while reading the story (Table 8). This pattern indicates that the intervention changed how participants engaged with narrative material, increasing their absorption into the narrated world and the extent to which the story captured their cognitive attention. This is informative because, as shown before, treated participants did not display greater baseline familiarity with writers' names; if anything, they scored lower on the author-recognition measure. The increase in these dimensions is therefore unlikely to reflect pre-existing differences in literary exposure and is instead consistent with an effect of the course itself.

Table 7:
Narrative engagement subscales

	Transportation		Realism		Cognitive access		Presence		Involvement		Loss of time		Loss of self		Distraction	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Treatment	0.293*	0.364**	0.082	0.145	0.247	0.336**	0.348**	0.432***	0.163	0.241**	0.234	0.341*	0.459***	0.523***	0.161	0.269
	(0.141)	(0.159)	(0.143)	(0.131)	(0.150)	(0.112)	(0.120)	(0.135)	(0.108)	(0.102)	(0.149)	(0.162)	(0.105)	(0.126)	(0.165)	(0.188)
Constant	4.081***	3.486***	4.863***	5.798***	4.958***	4.409***	4.829***	4.026***	5.210***	3.529***	4.237***	2.162**	4.469***	3.255**	5.023***	1.115
	(0.048)	(0.523)	(0.100)	(0.860)	(0.100)	(0.636)	(0.097)	(0.611)	(0.091)	(0.652)	(0.118)	(0.924)	(0.078)	(1.047)	(0.132)	(0.624)
Placebo mean	4.074	4.074	4.860	4.860	4.961	4.961	4.818	4.818	5.195	5.195	4.229	4.229	4.463	4.463	5.030	5.030
Observations	163	160	163	160	163	160	163	160	163	160	163	160	163	160	163	160
R^2	0.066	0.128	0.004	0.038	0.017	0.061	0.083	0.141	0.127	0.193	0.028	0.119	0.043	0.071	0.016	0.148
Wave fixed effects	✓		✓		✓		✓		✓		✓		✓		✓	
Demographic controls		✓		✓		✓		✓		✓		✓		✓		✓

Notes: Entries report coefficients from linear regressions of narrative engagement subscales on an indicator for assignment to the treatment condition. All specifications include fixed effects for the first experimental wave. Columns labeled *With controls* additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. The dependent variables are constructed as means of items from the Narrative Engagement and Immersion instrument (1–7 Likert scale): *Transportation* = mean of P41–P47, *Realism* = mean of P37–P40, *Cognitive access* = mean of P34–P36, *Presence* = mean of P21–P24, *Involvement* = mean of P25–P30, *Loss of time* = mean of P15–P17, *Loss of self* = mean of P18–P20, and *Distraction* = mean of P31–P33. See Appendix Table A.5 for the full list of items and coding details. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8:
Narrative engagement subscales: dichotomized measures

	Transportation		Realism		Cognitive access		Presence		Involvement		Loss of time		Loss of self		Distraction	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Treatment	0.127	0.150	0.062	0.073	0.039	0.074	0.065	0.080*	0.011	0.028	0.108*	0.129**	0.111*	0.134**	0.067	0.100**
	(0.078)	(0.084)	(0.063)	(0.065)	(0.060)	(0.044)	(0.046)	(0.039)	(0.049)	(0.040)	(0.050)	(0.053)	(0.052)	(0.045)	(0.048)	(0.043)
Constant	0.522***	-0.009	0.740***	1.034***	0.777***	0.656***	0.794***	0.104	0.859***	0.328	0.520***	-0.319	0.586***	0.406	0.713***	-0.311
	(0.037)	(0.256)	(0.047)	(0.129)	(0.033)	(0.194)	(0.043)	(0.244)	(0.040)	(0.284)	(0.041)	(0.407)	(0.034)	(0.376)	(0.036)	(0.340)
Placebo mean	0.519	0.519	0.740	0.740	0.779	0.779	0.792	0.792	0.857	0.857	0.519	0.519	0.584	0.584	0.714	0.714
Observations	163	160	163	160	163	160	163	160	163	160	163	160	163	160	163	160
R^2	0.033	0.102	0.006	0.032	0.019	0.067	0.022	0.095	0.017	0.068	0.012	0.080	0.019	0.047	0.008	0.116
Wave fixed effects	✓		✓		✓		✓		✓		✓		✓		✓	
Demographic controls		✓		✓		✓		✓		✓		✓		✓		✓

Notes: Entries report coefficients from linear regressions of an indicator equal to one if the corresponding narrative engagement subscale is above the midpoint of the 1–7 scale on an indicator for assignment to the treatment condition. All specifications include fixed effects for the first experimental wave. Columns with demographic controls additionally include age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. The dependent variables are based on the same subscales from the Narrative Engagement and Immersion instrument used in the continuous specifications: *Transportation* = mean of P41–P47, *Realism* = mean of P37–P40, *Cognitive access* = mean of P34–P36, *Presence* = mean of P21–P24, *Involvement* = mean of P25–P30, *Loss of time* = mean of P15–P17, *Loss of self* = mean of P18–P20, and *Distraction* = mean of P31–P33. See Appendix Table A.5 for the full list of items and coding details. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Ordered-logit estimates for individual Transportation items

	Coefficient		p-value		Odds ratio		ME: $\Pr(Y \geq 5)$	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
P41	0.519	0.704	0.001	0.000	1.680	2.021	0.052	0.071
P42	0.537	0.620	0.009	0.001	1.710	1.860	0.129	0.140
P43 (R)	0.106	0.217	0.655	0.385	1.112	1.243	0.026	0.051
P44	-0.087	0.047	0.678	0.832	0.917	1.048	-0.020	0.010
P45	0.605	0.715	0.035	0.067	1.832	2.045	0.084	0.093
P46	0.551	0.569	0.033	0.094	1.735	1.767	0.053	0.053
P47 (R)	0.307	0.462	0.192	0.115	1.360	1.587	0.066	0.093
Wave fixed effects	✓		✓		✓		✓	
Demographic controls		✓		✓		✓		✓

Notes: Entries report ordered-logit estimates for individual Transportation items from the Narrative Engagement and Immersion instrument. The coefficient corresponds to the treatment indicator. Odds ratios are obtained by exponentiating the ordered-logit coefficient. Marginal effects report the change in the predicted probability of choosing a response in the agreement range, defined as $Y \geq 5$ on the 1–7 Likert scale. See Appendix Table A.5 for the full list of items. Items P43 and P47 use their reverse-coded versions, so that higher values consistently indicate greater transportation. Baseline specifications include fixed effects for the first experimental wave. Specifications with demographic controls additionally include age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors are clustered at the treatment session term level.

4.2.2 Social-cognitive channels

Theory of Mind. To measure affective theory of mind from minimal cues, we administered the revised Reading the Mind in the Eyes Test (Baron-Cohen et al., 2001). Participants select which of four mental-state terms best matches the expression conveyed by the eye region in each item. Responses are scored as correct/incorrect and summed, so higher values indicate higher RMET performance Table 10 reports these results. We do not find robust evidence that the fiction-reading course improved RMET performance. While the fiction intervention affected how participants engaged with narrative material, it does not appear to have increased their ability to infer mental states from the eye region alone. In that sense, our results do not support a theory-of-mind interpretation of the treatment effect, at least as captured by this standard RMET measure.

Attributional complexity. We measured individual differences in the tendency to form multifaceted explanations of human behavior using the Attributional Complexity Scale (Fletcher et al., 1986). The scale comprises 28 statements answered on a bipolar agreement

Table 10: RMET index

	(1)	(2)
Treatment	0.190	0.239
	(0.237)	(0.247)
Constant	6.413***	5.548***
	(0.207)	(0.807)
Placebo mean	6.443	6.443
Observations	171	168
R^2	0.147	0.200
Wave fixed effects	✓	✓
Controls		✓

Notes: Entries report coefficients from linear regressions of performance on the Reading the Mind in the Eyes Test (RMET) on an indicator for assignment to the fiction-reading treatment. The dependent variable is an index equal to the total number of correctly identified mental states across RMET items (Baron-Cohen et al., 2001). Responses are coded as 1 if correct and 0 otherwise, and then summed, so that higher values indicate better affective theory-of-mind performance. In each RMET item, participants are shown a photograph of the eye region of a face and asked to identify the emotion or mental state that best describes the expression. The index therefore captures accuracy in inferring others' mental states from limited facial cues. See Appendix Figure A.4 for an example of how the question was asked. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

scale; items are combined so that higher values indicate greater attributional complexity. The scale is comprised of seven attribution subscales: *Complex*, *Situational*, *Motivation*, *Metacognitive*, *Ambiguity*, *Cognitive effort*, and *Behavior*. Table 11 presents the results for the attribution indices. The treatment effects here are more aligned with the idea that fiction changes how participants represent others in the first place. Relative to the placebo group, treated participants are more likely to endorse multifaceted explanations of human behavior: they appear more aware of the interdependencies and causes underlying others' behavior, more at ease with the ambiguity of others' motivations, and more likely to recognize that understanding another person requires effort. In short, others become less behaviorally reducible to simple or default interpretations. This pattern is consistent with fiction exposure fostering a richer and less categorical representation of other people, rather than simply increasing felt closeness or empathic concern.

Table 11:
Attribution indices

	Attribution (mean)		Complex		Situational		Motivation		Metacognitive		Ambiguity		Cognitive effort		Behavior	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Treatment	0.151	0.167*	0.286	0.291	0.118	0.109	0.031	0.032	0.169	0.222**	0.159*	0.190**	0.124	0.165**	0.169	0.158
	(0.091)	(0.091)	(0.180)	(0.177)	(0.138)	(0.132)	(0.105)	(0.092)	(0.093)	(0.090)	(0.084)	(0.079)	(0.076)	(0.056)	(0.116)	(0.125)
Constant	5.023***	4.841***	4.992***	4.406***	4.862***	5.067***	5.413***	4.582***	4.982***	4.371***	4.894***	5.076***	5.211***	5.664***	4.808***	4.719***
	(0.037)	(0.548)	(0.078)	(1.015)	(0.057)	(0.730)	(0.040)	(0.788)	(0.053)	(0.636)	(0.049)	(0.715)	(0.056)	(0.443)	(0.047)	(0.603)
Placebo mean	5.028	5.028	5.000	5.000	4.867	4.867	5.421	5.421	4.987	4.987	4.896	4.896	5.215	5.215	4.810	4.810
Observations	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169
R^2	0.028	0.096	0.034	0.082	0.015	0.037	0.025	0.072	0.018	0.119	0.010	0.057	0.015	0.102	0.010	0.098
First-wave fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Demographic controls		✓		✓		✓		✓		✓		✓		✓		✓

Notes: Entries report coefficients from linear regressions of attribution outcomes on an indicator for assignment to the treatment condition. The dependent variable in Columns (1)–(2) is the overall Attribution index, constructed as the respondent-level mean across 28 attribution items. Negatively keyed items (1, 2, 5, 6, 8, 11, 13, 16, 17, 18, 22, 23, 26, and 28) are reverse-coded as $8-x$, so that higher values consistently indicate stronger endorsement of complex, reflective, and context-sensitive attributional thinking. Columns (3)–(16) report results for the seven attribution subscales: *Complex*, *Situational*, *Motivation*, *Metacognitive*, *Ambiguity*, *Cognitive effort*, and *Behavior*. Each subscale is constructed as the respondent-level mean of four items following the questionnaire's block structure: *Complex* (items 1, 8, 15, and 22), *Situational* (items 2, 9, 16, and 23), *Motivation* (items 3, 10, 17, and 24), *Metacognitive* (items 4, 11, 18, and 25), *Ambiguity* (items 5, 12, 19, and 26), *Cognitive effort* (items 6, 13, 20, and 27), and *Behavior* (items 7, 14, 21, and 28). Thus, the subscales capture distinct dimensions of attributional reasoning, including preference for complex explanations, attention to situational influences, motivation to analyze behavior, metacognitive reflection, tolerance for ambiguity, cognitive effort, and behavioral reflection. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. See Appendix Table A.7 for the full list of items and coding details.

Narrative-based empathy. Table 12 reports the results for the empathy indices. In contrast to the evidence on narrative immersion and attributional complexity, we do not find robust treatment effects on either affective or cognitive empathy. This result is useful for interpreting the behavioral findings. If the fiction-reading course increased prosociality primarily by making participants feel more of what others feel, or by directly increasing standard empathy measures, we would expect movement in these indices. We do not observe such a pattern. Instead, the evidence is more consistent with the intervention affecting how participants imaginatively engage with stories and how they cognitively represent the complexity and individuality of others, rather than increasing empathy per se.

Table 12: Empathy indices

	Affective		Cognitive		Sympathy	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.013	0.031	-0.048	-0.018	0.052	0.046
	(0.098)	(0.101)	(0.107)	(0.127)	(0.098)	(0.088)
Constant	5.215***	4.042***	5.331***	5.834***	4.684***	3.818***
	(0.089)	(0.786)	(0.064)	(0.792)	(0.054)	(0.581)
Control mean	5.200	5.200	5.320	5.320	4.667	4.667
Observations	163	160	163	160	163	160
R^2	0.130	0.180	0.071	0.102	0.250	0.294
Wave fixed effects	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓

Notes: Entries report coefficients from linear regressions of attribution outcomes on an indicator for assignment to the treatment condition. The dependent variable in Columns (1)–(2) is the overall Attribution index, constructed as the respondent-level mean across 28 attribution items. Negatively keyed items (1, 2, 5, 6, 8, 11, 13, 16, 17, 18, 22, 23, 26, and 28) are reverse-coded as $8 - x$, so that higher values consistently indicate stronger endorsement of complex, reflective, and context-sensitive attributional thinking. Columns (3)–(16) report results for the seven attribution subscales: *Complex*, *Situational*, *Motivation*, *Metacognitive*, *Ambiguity*, *Cognitive effort*, and *Behavior*. Each subscale is constructed as the respondent-level mean of four items following the questionnaire's block structure: *Complex* (items 1, 8, 15, and 22), *Situational* (items 2, 9, 16, and 23), *Motivation* (items 3, 10, 17, and 24), *Metacognitive* (items 4, 11, 18, and 25), *Ambiguity* (items 5, 12, 19, and 26), *Cognitive effort* (items 6, 13, 20, and 27), and *Behavior* (items 7, 14, 21, and 28). Thus, the subscales capture distinct dimensions of attributional reasoning, including preference for complex explanations, attention to situational influences, motivation to analyze behavior, metacognitive reflection, tolerance for ambiguity, cognitive effort, and behavioral reflection. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. See Appendix Table A.7 for the full list of items and coding details.

4.2.3 Self-other representation channels

Self-other overlap. One possible mechanism is that fiction reading increases prosociality by making others feel closer to the self. To measure perceived closeness, we administered the Inclusion of Other in the Self (IOS) pictorial scale (Aron et al., 1992). Participants reported perceived closeness to a set of social targets (e.g., family, close friends, strangers, classmates, neighbors, and people with different socioeconomic or political backgrounds) by choosing among progressively overlapping self-other circles, with (1) being the farthest and (7) the closest. The results in Table 13 do not support this interpretation. Rather than increasing perceived closeness across the board, the treatment reduces self-other overlap for some specific groups, in particular study partners, neighbors, and individuals from lower socioeconomic strata. By contrast, we do not find significant changes in perceived overlap with family members, close friends, individuals from higher socioeconomic strata, people with different political views, people from other countries, or unknown others.

This pattern is interesting for two reasons. First, it is inconsistent with the idea that the fiction-reading intervention operated by making others feel more similar to the self in a generalized way. Second, the absence of effects for family and close friends on the one hand, and for more abstract or stereotyped categories such as foreigners, politically different others, and unknown people on the other, is also revealing. For familiar groups such as family and close friends, baseline overlap is already high (as can be seen by the placebo mean), leaving limited room for the treatment to increase it further. For more distant or abstract categories, responses may rely more on stable priors or categorical representations that are less easily shifted by the intervention. The treatment effects thus appear concentrated in intermediate social targets: groups that are socially meaningful and concrete enough to be represented as real others, but not so close that overlap is already maximal nor so abstract that responses are mainly driven by stereotypes.

Taken together, these results do not support a "closeness" account of the behavioral effects. If anything, they suggest that fiction may increase prosociality while making participants less likely to assimilate some others into the self. This is more consistent with an individuation-based interpretation, in which others become more psychologically distinct and less reducible to the self, rather than more similar to it.[10]

[10] We measured perceived social isolation using the UCLA Loneliness Scale (Version 3) (Russell, 1996). The instrument consists of 20 items answered on a 1–4 frequency scale ("never" to "always"), with positively worded items reverse-coded so that higher scores indicate greater loneliness. We don't uncover a pattern of increased social isolation, which suggests that the self-other overlap reduction is not necessarily accompanied with a generalized feeling of loneliness, see Table A.19 in the appendix.

Table 13:
Inclusion of the Other in the Self

	IOS Index (mean)		Family		Close friends		Study partners		Neighbors		Higher strata		Lower strata		Different political ideas		Other nationality		Unknown person	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Treatment	-0.323*	-0.329*	-0.068	-0.065	-0.146	-0.196	-0.532**	-0.545**	-0.670***	-0.606***	-0.341	-0.267	-0.573**	-0.607**	-0.396*	-0.389	-0.126	-0.153	-0.054	-0.137
	(0.156)	(0.164)	(0.200)	(0.194)	(0.198)	(0.196)	(0.173)	(0.174)	(0.207)	(0.188)	(0.237)	(0.254)	(0.222)	(0.240)	(0.213)	(0.223)	(0.134)	(0.126)	(0.170)	(0.190)
Constant	3.590***	5.276***	5.269***	8.137***	5.137***	7.360***	4.113***	5.767***	3.019***	4.737***	3.423***	4.867***	3.682***	5.853***	2.621***	4.477***	3.243***	3.494***	1.802***	2.787***
	(0.118)	(0.736)	(0.114)	(1.042)	(0.142)	(0.952)	(0.161)	(0.721)	(0.168)	(1.107)	(0.202)	(1.363)	(0.120)	(1.011)	(0.168)	(1.204)	(0.104)	(0.906)	(0.136)	(0.726)
Placebo mean	3.586	3.586	5.266	5.266	5.127	5.127	4.114	4.114	3.013	3.013	3.418	3.418	3.684	3.684	2.620	2.620	3.241	3.241	1.797	1.797
Observations	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169
R^2	0.035	0.099	0.002	0.050	0.019	0.057	0.035	0.076	0.062	0.082	0.015	0.102	0.039	0.090	0.024	0.088	0.003	0.081	0.009	0.065
First-wave fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Demographic controls		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓

Notes: Entries report coefficients from linear regressions of IOS (Inclusion of the Other in the Self) outcomes on an indicator for assignment to the treatment condition. The IOS instrument asks respondents to choose, for each target group, the pair of overlapping circles that best represents their perceived closeness to that group, with higher values indicating greater perceived overlap between self and other. The dependent variable in Columns (1)–(2) is the IOS index, constructed as the respondent-level mean across the nine target-specific IOS items. The remaining columns report results separately for each IOS item: *Family* (Columns 3–4, item 1), *Close friends* (Columns 5–6, item 2), *Study partners* (Columns 7–8, item 9), *Neighbors* (Columns 9–10, item 8), *Higher strata* (Columns 11–12, item 5), *Lower strata* (Columns 13–14, item 6), *Different political ideas* (Columns 15–16, item 4), *Other nationality* (Columns 17–18, item 7), and *Unknown person* (Columns 19–20, item 3). See Appendix Table A.8 for the full list of items and coding details. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Individuation. To capture an *individuation* orientation—the tendency to perceive others as distinctive individuals rather than interchangeable category members—we administered a 10-item Perceived Uniqueness scale developed for this study, informed by the individuation-versus-categorization framework (Fiske & Neuberg, 1990). The items capture endorsement of people's unique qualities, attention to differentiating details, reliance on context when forming impressions, and resistance to stereotype or category-based thinking. We pre-specified two measures: (i) a broad *Uniqueness Orientation* index (mean of all items, after reverse-coding), and (ii) an *Attention to Individual Differences* subscale (two items) that isolates explicit attentiveness to what makes one person different from another. Appendix Table A.3 reports full item wording and coding. For each of the 10 items, participants rated their agreement with each statement on a 1–5 Likert scale (from "strongly disagree" to "strongly agree"). The items are designed to measure a broad orientation toward individuating others, that is, attending to idiosyncratic attributes and contexts, avoiding categorical thinking, and rejecting the view that people are interchangeable. We reverse-coded negatively keyed items so that higher values always indicate greater individuation.

We construct two indices from these items. First, the Uniqueness Orientation index is the respondent-level mean of all ten items (Cronbach's $\alpha = 0.77$), intended to capture a general tendency to perceive and evaluate others as psychologically distinct individuals rather than as interchangeable members of categories. This broader index combines several related dispositions: recognizing that people have unique qualities, paying attention to what makes a person different, incorporating contextual information when forming impressions, resisting stereotype- or category-based judgments, and rejecting the idea that people are primarily defined by the groups to which they belong. Higher values indicate a stronger tendency to represent others in individualized rather than categorical terms; lower values indicate greater comfort with interchangeable or group-based representations.

Second, we construct an Attention to Individual Differences index as a two-item subscale (Cronbach's $\alpha = 0.72$). This narrower measure isolates explicit attention to differentiating features across people: endorsing that one notices what makes someone different and rejecting the statement that one does not pay attention to the details that differentiate individuals. Higher values indicate greater attentiveness to fine-grained interpersonal differences; lower values indicate less attention to what distinguishes one person from another.

In the pooled sample (Table 14), we do not find a statistically significant effect of the fiction course on the broader Uniqueness Orientation index. However, the Attention to Individual Differences index shows a positive treatment effect that is significant at conventional levels. When we disaggregate by session timing (immediate post-course versus four-week follow-up), the observed effects appear to weaken over time. In particular, the treatment effect on Attention to Individual Differences is concentrated in the immediate post-course measurement ($p = 0.007$) and is no longer present in the follow-up. This pattern suggests that the intervention increased participants' attentiveness to what differentiates one person from another, consistent with the idea that fiction makes others less reducible to generic or category-based representations.

Table 14: Perceived Uniqueness Indices

	Uniqueness Orientation		Attention to Individual Differences	
	(1)	(2)	(3)	(4)
Treatment	0.033	0.042	0.251**	0.259**
	(0.063)	(0.067)	(0.111)	(0.103)
Constant	3.906***	3.560***	3.546***	3.864***
	(0.044)	(0.406)	(0.060)	(0.381)
Placebo mean	3.905	3.905	3.545	3.545
Observations	163	160	163	160
R^2	0.002	0.019	0.018	0.059
Placebo mean	3.545	3.545	2.193	2.193
Observations	163	160	163	160
R^2	0.002	0.021	0.018	0.055
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of perceived uniqueness (individuation) measures on an indicator for assignment to the fiction-reading treatment. The dependent variables are two pre-specified indices constructed from the Perceived Uniqueness scale developed for this study. The *Uniqueness Orientation* index is defined as the mean of ten items capturing endorsement of individual distinctiveness, attention to differentiating features, contextualized judgment, and resistance to category-based thinking. Negatively worded items (Items 3, 6, and 10) are reverse-coded so that higher values consistently indicate greater individuation. The *Attention to Individual Differences* subscale is defined as the mean of two items explicitly capturing attentiveness to what differentiates one person from another (Items 2 and reverse-coded Item 3). See Appendix Table A.3 for the full list of items and coding details. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.3 Robustness

No increase in moral self-concept. One possible concern is that the behavioral effects reflect participants viewing themselves as more moral after the intervention, rather than a change in how they weigh others' welfare in decision-making. To measure the extent to which moral traits are central to the self-concept, we administered the Moral Identity Scale (Aquino & Reed, 2002). Following the standard structure of the instrument, participants evaluate a set of moral traits and then respond to items capturing the centrality of these traits, typically separating symbolization from internalization; higher values indicate a more

central moral identity. The evidence does not support this interpretation. As shown in Table A.20 we find no robust treatment effect on moral self-concept. This helps narrow the interpretation of the main findings: the increase in generosity and trustworthiness does not appear to be driven by participants seeing themselves as more moral individuals.

No differential attrition across treatment cells. A natural concern is that the estimated treatment effects may be driven by selective participation in the experimental sessions rather than by the fiction-reading intervention itself. Appendix A.3.5 addresses this possibility directly. We find no evidence of differential attrition across treatment cells, either overall or conditional on baseline characteristics. This reduces the concern that the main results are driven by compositional differences between the observed treatment and placebo samples.

No evidence that the effects are short-lived. A second concern is that the treatment effects may simply reflect short-run priming immediately after the intervention. To assess this possibility, we compare outcomes measured immediately after the course with those measured in the four-week follow-up. We do not find evidence that the main behavioral effects are short-lived. The broad pattern of higher altruism and greater trustworthiness remains similar across both horizons, even if some individual coefficients vary in precision.

For the psychological scales, the evidence is also broadly stable across the two measurement windows. We do not find systematic differences between the short- and longer-run follow-up for most of these measures. An exception is the narrower individuation measure, where the effect appears more concentrated immediately after the course. This suggests that some changes in how participants attend to others may be strongest in the short run, even if the behavioral effects do not disappear over the same horizon.

The placebo does not appear to reduce prosociality. Another possible concern is that the estimated treatment effects may partly reflect a detrimental effect of the placebo condition, rather than a positive effect of the fiction-reading course. To explore this possibility, we compare behavior and psychological outcomes within the placebo arm by degree of exposure, distinguishing participants who attended at least three sessions from those with lower participation. We do not find evidence that greater exposure to the placebo course is associated with lower prosocial behavior or systematically less prosocial psychological profiles. This makes it less likely that the main treatment effects are driven by the placebo condition reducing prosociality.

Taken together, these robustness checks reinforce the main interpretation of the paper. The increase in altruism and trustworthiness is unlikely to be explained by selective attrition, by a temporary post-intervention effect, by changes in moral self-perception, or by the placebo condition having a negative impact on prosociality.

5 Discussion

Our findings provide causal evidence that guided exposure to literary fiction can increase prosocial behavior when measured through revealed preferences rather than self-reports alone: fiction does not merely change what participants say about how they would treat others, but how they act when prosociality is costly. Relative to a placebo course, participants in the fiction condition give more money in the dictator game and behave more trustworthily in the trust game. Our incentive-compatible changes in behavior speak to the revealed-preference approach emphasized by Dodell-Fedder and Tamir, 2018. We should also emphasize that we are finding that the effects of fiction in prosociality are indeed causal, and not caused by selection as some have recently argued (Dubourg & Baumard, 2023). Our paper supports the most classical view that reading *causally* makes people more prosocial.

The evidence also points to a mechanism centered less on empathy, moral self-concept or self-other overlap, and more on a gain in social-cognitive complexity regarding how participants see others. Treated participants show more narrative transportation, more attributional complexity, and greater attention to what differentiates one person from another. Taken together, this suggests that fiction works by moving readers away from simplified, category-based representation of other people and toward richer models in which others are seen as psychologically textured individuals—with motives, ambiguities, histories, and contradictions that deserve attention. This notion is supported by an important finding of our results: reading causally decreases closeness to others (measured by the Inclusion of Other in the Self scale) only when they are initially perceived as far. It is not that reading makes us closer to others, it helps us understand how complex others are.

A further implication of our results is that prosociality can increase without evidence of tighter social norms. We do not find evidence of changes in empirical or normative expectations. If anything, in some games participants become slightly more permissive in their personal and normative judgments of behavior that departs from conventional prosocial expectations. This is promising in terms of scalable interventions with the goal of increasing prosociality. We know that social norm change is difficult and costly, as it requires coordination and strategic choices. Our results suggest that this social norm change is not necessary for an increase in prosociality, which can be increased by an individually-targeted intervention around reading.

More broadly, these findings help clarify what kind of educational intervention a guided fiction-reading course may be. Its value may lie not in teaching a norm directly, or boosting morality or empathy but in weakening reductive models of other people. In that sense, literary reading may create welfare gains not by telling readers what is right, but by changing the kind of social world they perceive when making choices about others.

References

Alan, S., Baysan, C., Gumren, M., & Kubilay, E. (2023). Belief in hard work and prosocial behavior. *Working Paper*.

Alesina, A., Stantcheva, S., & Teso, E. (2018). Intergenerational mobility and preferences for redistribution. *American Economic Review*, *108*(2), 521–554.

Alexander Cappelen, A. S., John List, & Tungodden, B. (2020). The effect of early-childhood education on social preferences. *Journal of Political Economý*, *128*(7). https://doi.org/https://doi.org/10.1086/706858

Andries, M., Bursztyn, L., Chaney, T., & Djourelova, M. (2024). In their shoes: Empathy through information. *Working Paper*.

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology*, *83*(6), 1423–1440. https://doi.org/10.1037/0022-3514.83.6.1423

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612. https://doi.org/10.1037/0022-3514.63.4.596

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

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Braun, R. (2022). Bloodlines: National border crossings and antisemitism in weimar germany. *American Sociological Review,*

Brodsky, J. (1987). Nobel lecture [Nobel Prize in Literature. Delivered December 8, 1987. Available at https://www.nobelprize.org/prizes/literature/1987/brodsky/lecture/].

Brodsky, J. (1991). The condition we call exile. *Culture, Theory and Critique*, *34*(1), 1–8.

Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology*, *12*(4), 321–347. https://doi.org/10.1080/15213260903287259

Dodell-Fedder, D., & Tamir, D. I. (2018). Reading has a small positive impact on social cognition: A meta-analysis. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000395.supp

Dodell-Feder, D., Liu, G., Amormino, P., Handley, E. D., & Tamir, D. (2025). The impact of fiction reading on social outcomes: A four-week randomized controlled study. *Psychology of Aesthetics, Creativity, and the Arts*.

Dubourg, E., & Baumard, N. (2023, January). Do fictions impact people's beliefs? a critical view. Taylor; Francis. https://doi.org/10.4324/9781003119456-14

Enke, B. (2025). Moral universalism and the structure of ideology. *American Economic Review*.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Figueroa, D., & Fouka, V. (2023). Structural transformation and value change: The british abolitionist movement. *NBER Working Paper*.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–74, Vol. 23). Academic Press. https://doi.org/10.1016/S0065-2601(08)60317-2

Fletcher, G. J. O., Danilovics, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual differences measure. *Journal of Personality and Social Psychology*, *51*(4), 875–884. https://doi.org/10.1037/0022-3514.51.4.875

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*(3), 347–369. https://doi.org/10.1006/game.1994.1021

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*. https://doi.org/https://doi.org/10.1162/003355300554917

Green, M. C., & Brock, T. C. (2000a). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*.

Green, M. C., & Brock, T. C. (2000b). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701–721. https://doi.org/10.1037/0022-3514.79.5.701

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Hartinger, K., Resnjanskij, S., Ruhose, J., & Wiederhold, S. (2026). Individualism, human capital formation, and labor market success. *Journal of the European Economic Association*.

Keen, S. (2007). *Empathy and the novel*. Oxford University Press.

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

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*, 495–524. https://doi.org/10.1111/jeea.12006

Leonardo Bursztyn, A. L. G., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American Economic Review*, *110*. https://doi.org/https://DOI:10.1257/aer.20180975

Nagel, T. (1970). *The possibility of altruism*. Clarendon Press.

Natalia Amaya, J. J. R., Allison Benson, & Hosie, R. S. (2024). Unveiling classism: The potential of experiential-based reflection in transforming perceptions. *Documento CEDE-TREE*. https://doi.org/10.1037/xge0000395.supp

Nussbaum, M. C. (1990). *Love's knowledge : Essays on philosophy and literature*. Oxford University Press.

Read, D., & van Leeuwen, B. (1998). Predicting hunger: The effects of appetite and delay on choice. *Organizational Behavior and Human Decision Processes*, *76*.

Russell, D. W. (1996). Ucla loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment*, *66*(1), 20–40. https://doi.org/10.1207/s15327752jpa6601_2

Stantcheva, S. (2022). Understanding inequality: Beliefs, perceptions, and narratives. *Economica*.

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

A Appendix

Initial registration: May 15, 2025	Update on November 10, 2025
Economic games	
Dictator game	Dictator game
Ultimatum game	Ultimatum game
Trust game	Trust game
Third party game	Third party game
Psychological Scales	
	Inclusion of the Other in the Self (IOS) scale
	Reading the Mind in the Eyes test (RMET)
	A questionnaire about the value given to the aesthetic experience of the other.

Table A.1: Preregistration under Title *Making inequality unacceptable: an experiment on fiction reading and prosociality.* RCT ID AEARCTR-0015996

A.1 Design materials

A.1.1 Recruitment

Estimado/a estudiante,

La **Facultad de Economía de la Universidad de los Andes** te invita a participar en un **curso sobre ciencias del comportamiento, literatura y aplicaciones en la vida diaria.**

¿En qué consiste?

- Asistir a **4 sesiones del curso** (cada una de aprox. 1 hora y 20 min).
- Participar en **2 actividades/encuestas** posteriores a las sesiones (1 hora y 20 min de duración cada una)

Reconocimiento económico 💰

- Recibirás **$80.000** si participas en las sesiones.
- Además, podrás obtener **pagos adicionales (de hasta $ 50.000)** por tu participación y decisiones en las actividades posteriores.

Lo que dicen quienes ya participaron 👇

Figure A.1: Invitation email

The following images summarize our local adaptation of the Author Recognition test. In Figure A.2 the instruction given to participants and in Figure A.3 the full set of names used.

20. En esta pregunta le vamos a proponer un juego (o quiz). A continuación encontrará una lista de nombres. Algunos de ellos son autores de libros de ficción, otras son personas reconocidas por otros motivos.

Su tarea es marcar cuáles corresponden a autores de al menos un libro de ficción que se haya publicado (o reeditado) en los últimos 20 años. Por cada nombre que usted marque correctamente obtendrá un punto, pero se le restará un punto por cada nombre que marque y no corresponda a un autor de ficción. Es importante que no se demore más de 10 minutos respondiendo, porque si no su puntaje no será contabilizado.

Figure A.2: Instructions given to participants for the ART in the Entrance Survey

Writers	Non-writers
Sergio Ramírez°°	Daron Acemoglu °°
Claudia Piñeiro°°	Mariana Mazzucatto °°
Antonio Muñoz Molina°°	Mauricio Cárdenas
Cristina Rivera Garza°°	Elinor Ostrom ** °°
Pilar Quintana	Paul Krugman °°
Fernanda Trías	Claudia Goldin °°
María Ospina Pizano	Joseph Stiglitz °°
Laura Ortiz Gómez	Esther Duflo °°
Lorena Salazar Masso	James A. Robinson°°
Laura Restrepo	Juan Camilo Cárdenas
Sara Jaramillo Klinkert	Marcela Melendez
Catalina Navas	Hernando Zuleta
Alba Lucía Ángel	Marcela Eslava
Juan Gabriel Vásquez	Francisco Gutierrez Sanín
C. S. Lewis°°	Cecilia López
J.K. Rowling°°	Mauricio García Villegas
Mario Mendoza	Cecilia Orozco
Piedad Bonett	Thierry Ways
Rosa Montero	Héctor Riveros
Pablo Montoya	Mauricio Reina
Fernando Vallejo	Natalia Ángel Cabo
William Ospina	Daniel Coronell
J. M. Coetzee°°	María Jimena Duzán
Mario Vargas Llosa°°	Rodrigo Upprimmy
Alice Munro°°	Ana Bejarano
Annie Ernaux°°	J.M. Johnson
Orhan Pamuk°°	Pedro Pablo Kuczynski
Marvel Moreno**	Justin Trudeau°°
Germán Espinosa**	Sophie Calle°°
Gilmer Mesa	Christine Lagarde°°
Ricardo Silva Romero	Ana Cristina Restrepo Jiménez
Alessandro Baricco°°	Claudia Palacios
John Banville°°	Claudia Sheinbaum
Arturo Pérez Reverte°°	Armando Montenegro
Álvaro Mutis**	Mario Draghi°°
Jose Eustasio Rivera**	Moisés Wasserman
Isabel Allende°°	Yolanda Ruiz
Hector Abad Faciolince	Jorge Espinosa
Elisa Mujica	Alberto Núñez Feijóo°°
Juan Esteban Constaín	Camila Vallejo°°
Jorge Franco	María Isabel Rueda
Jorge Volpi°°	Sandra Borda
Margarita García Robayo	Lucas Pombo
Legend	
**Dead	10
°° Non-colombian	30

Figure A.3: Full set of names used in the ART

A.1.2 Fiction Reading course material

Session 1 Foundations of Literary Reading. The course begins by reframing literary reading as a practice of disciplined attention rather than content consumption. Students are introduced to narrative perspective, voice, focalization, and the distinction between plot ("what happens") and form ("how it is told"). The session trains readers to attend to how texts construct access to characters' inner worlds, encouraging slower, more reflective engagement with psychological detail.

Session 2 Mental Simulation and Theory of Mind. This session focuses on fiction as a tool for mental simulation. Students analyze how narratives invite readers to inhabit characters' mental states, infer intentions, and navigate ambiguity. The emphasis is on theory-of-mind processes: tracking motivations, conflicting desires, unreliable narration, and partial information. Literary form is explicitly linked to the cognitive practice of modelling others' interiority.

Session 3 Emotional Reflection and Narrative Engagement. Students explore the emotional dimension of literary experience. Through discussion of short stories, images, and narrative passages, the session examines how fiction elicits empathy, discomfort, ambiguity, and moral tension. The goal is not emotional persuasion but emotional differentiation—learning to recognize complex, mixed, and sometimes contradictory responses to characters and situations.

Session 4 Reading, Well-Being, and Reflective Practice. The final session connects literary engagement to well-being and self-regulation. Students discuss how sustained attention to narrative complexity may reduce cognitive rigidity and encourage reflective judgment. The session reinforces habits of close reading, ambiguity tolerance, and individuated attention—linking literary practice to everyday social perception.

A.1.3 Placebo course: Behavioral science

Session 1 Decision Theory and Behavioral Biases. The course opens with core ideas from economic decision theory and behavioral economics. Students learn how preferences, beliefs, and constraints shape choices, and how systematic biases (context effects, heuristics, framing) influence behavior. The session introduces practical applications such as nudges and incentive design.

Session 2 The Planning Fallacy and Time Management. This session examines why individuals systematically underestimate project duration. Mechanisms such as optimism

bias, projection bias, and neglect of intermediate steps are discussed. Students are introduced to the "outside view" as a corrective tool, emphasizing structured forecasting rather than introspective simulation.

Session 3 Incentives, Intrinsic Motivation, and Social Image. Students analyze classic experiments on incentive crowding-out and distrust in cooperation. The session discusses how monetary incentives interact with intrinsic motivation, reciprocity, and social image concerns. It also examines limits of perspective-taking and highlights the importance of eliciting others' views ("perspective-getting") instead of relying on introspection.

Session 4 — Social Norms and Collective Behavior. The final session introduces empirical and theoretical frameworks for studying social norms. Students distinguish descriptive and injunctive norms, examine pluralistic ignorance, and discuss how expectations shape cooperation and conformity. The focus is conceptual and analytical rather than experiential or narrative.

A.1.4 Psychological scales: description and scoring

This Appendix summarizes the psychological instruments used to probe mechanisms potentially linking fiction exposure to prosociality. Unless otherwise noted, we construct scale scores by reverse-coding negatively keyed items, averaging (or summing) items within a scale, and then standardizing for analysis.

Perceived Uniqueness / Individuation (study-developed scale) Table A.3 below provides the 10-item questionnaire designed to measure individual differences in the tendency to perceive others as distinctive individuals rather than interchangeable category members. Items are answered on a 1–5 Likert scale. We reverse-code items 3, 6, and 10 as $x_r = 6 - x$ so that higher values consistently indicate stronger individuation.

Loneliness We measure perceived social isolation using the UCLA Loneliness Scale (Version 3) (Russell, 1996). The instrument comprises 20 items answered on a 1–4 frequency scale (Table A.4). Positively worded items are reverse-coded and responses are summed (or averaged) so that higher scores indicate greater loneliness. Items are answered on a 1–4 frequency scale: (1) I often feel this way; (2) I sometimes feel this way; (3) I rarely feel this way; (4) I never feel this way.

Table A.2: Psychological instruments

Scale	**# Items**	**Response Format**	**Index Construction**	**Cronbach's α**
Uniqueness	10	1–5 Likert	Overall mean	0.77
Uniqueness attention component	2	1–5 Likert	Subscale mean	0.72
UCLA Loneliness	20	1–4 frequency	Overall mean	0.93
Narrative Engagement and Immersion	47	1–7 Likert	Subscale means	
Affective	7	1–7 Likert	Mean of P1–P6, P8	0.83
Cognitive	6	1–7 Likert	Mean of P9–P14	0.81
Sympathy	4	1–7 Likert	Mean of P6-P8	0.50
Loss of time	3	1–7 Likert	Mean of P15–P17	0.67
Loss of selfawarnes	3	1–7 Likert	Mean of P18–P20	0.66
Presence	4	1–7 Likert	Mean of P21–P24	0.72
Involment	6	1–7 Likert	Mean of P25–P30	0.86
Distraction	3	1–7 Likert	Mean of P31–P33	0.87
Cognitive access	3	1–7 Likert	Mean of P34–P36	0.66
Realism	4	1–7 Likert	Mean of P37–P40	0.80
Transportation	7	1–7 Likert	Mean of P41–P47	0.65
Moral Identity	13	1–7 Likert	Internalization / Symbolization means	
Internalization	5	1–7 Likert	Mean of P1, P2, P4, P5 and P7	0.60
Symbolization	5	1–7 Likert	Mean of P8–P13	0.75
Attributional	28	1–7 Likert	Mean	0.89
Complexity	4	1–7 Likert	Mean of P1, P8, P15 and P22	0.77
Situacional	4	1–7 Likert	Mean of P2, P9, P16 and P23	0.55
Motivation	4	1–7 Likert	Mean of P3, P10, P17 and P24	0.63
Metacognitive	4	1–7 Likert	Mean of P4, P11, P18 and P25	0.63
Ambiguity	4	1–7 Likert	Mean of P5, P12, P19 and P26	0.49
Inclusion of Other in Self (IOS)	9	7-option pictorial	Target-specific score	–
Reading the Mind in the Eyes (RMET)	36	4-option multiple choice	Sum correct responses	–

Notes: Cronbach's α is reported for multi-item self-report indices. It is not defined for single-item measures (IOS) or for performance-based tests scored by correctness (RMET). For Narrative Experience and Moral Identity, we report the overall index α and the range of α for subscales.

#	Item	R?

Table A.4: Loneliness items

#	Item	R?
P1	How often do you feel "in tune" with the people around you?	R
P2	How often do you feel that you lack companionship?	
P3	How often do you feel that you have no one to turn to?	
P4	How often do you feel alone?	
P5	How often do you feel part of a group of friends?	R
P6	How often do you feel that you have a lot in common with the people around you?	R
P7	How often do you feel that you are no longer close to anyone?	
P8	How often do you feel that the people around you do not share your interests and ideas?	
P9	How often do you feel outgoing and friendly?	R
P10	How often do you feel close to other people?	R
P11	How often do you feel left out?	
P12	How often do you feel that your relationships with other people are not meaningful?	
P13	How often do you feel that no one really knows you well?	
P14	How often do you feel isolated from others?	
P15	How often do you feel that you can find companionship when you want it?	R
P16	How often do you feel that there are people who really understand you?	R
P17	How often do you feel shy?	
P18	How often do you feel that people are around you but not with you?	
P19	How often do you feel that there are people you can talk to?	R
P20	How often do you feel that there are people you can turn to?	R
–	*Attention:* "To assess your level of attention, please select the option 'I rarely feel this way' for this question."	–

Narrative engagement and immersion (post-story battery) Following the reading of *El hijo*, participants completed a 47-item narrative-experience battery capturing emotional engagement, perspective-taking within the story, attentional focus, immersion, narrative presence, and comprehension (see Table A.5). The instrument draws on the Transportation framework of Green and Brock (2000b) and the multidimensional Narrative Engagement

Table A.3: Perceived Uniqueness / Individuation questionnaire (items and coding)

#	Item (English translation)	R?
P1	Each person has unique qualities; they are not just a label.	
P2	When I meet someone, I notice what sets them apart from others.[a]	
P3	I do not usually pay attention to details that distinguish one person from another.	**R**
P4	To form an opinion about someone, it matters to know their story or context.	
P5	I make an effort not to rely on stereotypes when thinking or talking about others.	
P6	It is easier for me to think of people as a group rather than as individuals.	**R**
P7	Even if they seem similar, each person retains their own identity.	
P8	I like discovering what makes someone special, even briefly.[a]	
P9	I treat each person as an individual, not as a social category.	
P10	Many people are not that different; they seem interchangeable.	**R**

Notes: Items were answered on a 1–5 Likert scale ("strongly disagree" to "strongly agree"). "R" denotes reverse-coded items (coded as $x_r = 6 - x$) so that higher values always indicate greater individuation / uniqueness orientation. We construct (i) a *Uniqueness Orientation* index as the mean of all 10 items after reverse-coding, and (ii) an *Attention to Individual Differences* index as the mean of the two items marked with [a] (items 2 and 8). This scale was developed for this study, drawing on the individuation-versus-categorization framework in social cognition (Fiske & Neuberg, 1990).
Attention check: We additionally included an item instructing respondents to mark "strongly agree" (not part of the indices).

model of Busselle and Bilandzic (2009), but expands upon these original scales. Items are answered on a 1–7 Likert scale. Negatively keyed items are reverse-coded. We construct sub-indices corresponding to emotional engagement, attentional focus, immersion, and comprehension, as well as an overall engagement index. Items are answered on a 1–7 Likert scale: Strongly disagree; Very much disagree; Disagree; Neither agree nor disagree; Agree; Very much agree; Strongly agree.

Table A.5: Narrative-experience items after reading *El hijo*

#	Item (English translation)	R?
P1	At key moments in the story, I felt that I knew exactly what the characters were experiencing emotionally.	
P2	At important moments in the story, I could feel the emotions the characters were feeling.	
P3	During the story, when a main character succeeded, I felt happy; and when they suffered in some way, I felt sad.	
P4	I never really shared the characters' emotions.	**R**

#	Item (English translation)	R?
P5	The story affected me emotionally.	
P6	I felt compassion for some of the characters in the story.	
P7	I felt embarrassed for some of the characters in the story.	
P8	I worried about some of the characters in the story.	
P9	I could understand the events in the story in a way similar to how the characters understood them.	
P10	I understood the reasons why the characters acted as they did.	
P11	I could understand why the characters felt the way they did.	
P12	My understanding of the characters is unclear.	**R**
P13	I found it difficult to understand why the characters reacted the way they did.	**R**
P14	I could easily imagine myself in the situation of some of the characters.	
P15	While reading the story, I lost track of time.	
P16	The story seemed slow to me.	**R**
P17	When the story ended, I was surprised that it had passed so quickly.	
P18	At some points in the story, I completely forgot that I was participating in an experiment.	
P19	I forgot my own problems and concerns while reading.	
P20	While reading the story, I found myself thinking about what I had done before or what I would do afterward.	**R**
P21	At some points in the story, the narrated world felt closer to me than the real world.	
P22	My attention was more on my surroundings than on the story.	**R**
P23	The story created a new world, and that world suddenly disappeared when the story ended.	
P24	While reading the story, my body was in the room, but my mind was inside the world created by the story.	
P25	I was mentally involved in the story while reading it.	
P26	I never really felt drawn into the story.	**R**
P27	While reading, I felt completely immersed in the story.	
P28	Overall, the experience of reading the story was intense for me.	
P29	I wanted to know how the story ended.	
P30	While reading, I wanted to know how events would unfold.	
P31	I found my mind wandering while reading the story.	**R**
P32	While reading the story, I caught myself thinking about other things.	**R**

#	Item (English translation)	R?
P33	I found it difficult to keep my mind on the story.	**R**
P34	I could easily follow the action and events being narrated.	
P35	I had difficulty recognizing the thread of the story.	**R**
P36	I had to make an effort to stay focused on the story.	**R**
P37	The story was logical and convincing.	
P38	I understood why the events unfolded the way they did.	
P39	At some points in the story, it was not clear to me why something was happening.	**R**
P40	At some points, I found it difficult to understand what was happening in the story.	**R**
P41	While reading the story, I could easily imagine the events taking place.	
P42	I could imagine myself in the scene of the narrated events.	
P43	After finishing the story, I found it easy to put it out of my mind.	**R**
P44	I found myself thinking about ways the story could have ended differently.	
P45	The events in the story are relevant to my everyday life.	
P46	The events in the story have changed my life.	
P47	While reading the story, I was aware of the activity taking place in the room around me.	**R**
–	*Attention:* "To assess your level of attention, please mark 'Neither agree nor disagree' on this question."	–

Affective and cognitive empathy (state, narrative-based) Although the survey did not include a dedicated trait-empathy instrument (e.g., the Interpersonal Reactivity Index), we construct exploratory state-based empathy indices from the post-story narrative battery based on Table A.5.

We define **Affective Empathy** as the extent of emotional sharing and compassion toward the story characters. This index is constructed as the mean of items P1, P2, P3, P4 (reverse-coded), P5, P6, and P8. Higher values indicate stronger emotional resonance with the characters.

We define **Cognitive Empathy** as perspective-taking and understanding of characters' mental states. This index is constructed as the mean of items P9–P14 (with P12 and P13 reverse-coded). Higher values indicate greater narrative-based perspective-taking.

These indices capture empathy within the narrative context and should not be interpreted as measures of general empathic disposition.

Moral identity (self-importance of morality) We measure the self-importance of moral traits using the Moral Identity Scale (Aquino & Reed, 2002). Participants are first presented with a set of moral traits and then respond to items capturing how central these traits are to their identity (see Table A.6). The standard scoring distinguishes internalization (private centrality) from symbolization (public expression), with higher values indicating a more central moral identity.

Items are answered on a 1–7 Likert scale (as above). The survey first presents nine moral traits (e.g., Cuidadoso/a, Compasivo/a, Justo/a, Amistoso/a, Generoso/a, Servicial, Trabajador/a, Honesto/a, Amable), then asks respondents to rate statements about how central these characteristics are to their identity.

Table A.6: Moral identity items shown in the survey

#	**Item (English translation)**	**R?**
P1	It would make me feel good to be a person who has these characteristics.	
P2	Being someone who has these characteristics is an important part of who I am.	
P3	A large part of my emotional well-being is tied to having these characteristics.	
P4	I would be ashamed to be a person who has these characteristics.	**R**
P5	Having these characteristics is not really important to me.	**R**
P6	Having these characteristics is an important part of my sense of self.	
P7	I have a strong desire to possess these characteristics.	
P8	I often buy products that communicate the fact that I have these characteristics.	
P9	I often wear clothes that identify me as someone who has these characteristics.	
P10	The kinds of things I do in my spare time (e.g., hobbies) clearly identify me as someone who has these characteristics.	
P11	The kinds of books and magazines I read identify me as someone who has these characteristics.	
P12	The fact that I have these characteristics is communicated to others by my membership in certain organizations.	
P13	I actively participate in activities that communicate to others that I have these characteristics.	
–	*Attention:* "To check your level of attention, please choose the option 'Strongly agree' for this question."	–

Attributional complexity We measure attributional complexity using the Attributional Complexity Scale (Fletcher et al., 1986). The instrument consists of 28 statements about

explaining human behavior, answered on a bipolar agreement scale (Table A.7). After accounting for item direction, responses are aggregated so that higher values reflect a greater tendency to construct multifactorial explanations.

Items are answered on a 1–7 Likert scale (as above).

Table A.7: Attributional complexity items shown in the survey

#	Item (English translation)	R?
P1	I am usually not interested in analyzing and explaining people's behavior.	**R**
P2	Once I find a single cause for a person's behavior, I usually do not look for additional explanations.	**R**
P3	I believe it is important to analyze and understand our own thought processes.	
P4	I think a lot about the influence I have on other people's behavior.	
P5	I have found that the relationships among a person's attitudes, beliefs, and character traits are usually simple and straightforward.	**R**
P6	If I see someone behaving in a very strange or unusual way, I usually attribute it to them being a strange or unusual person and do not worry about explaining it further.	**R**
P7	I have reflected quite a bit on the family background and personal history of people close to me in order to understand why they are the kind of people they are.	
P8	I do not enjoy getting into discussions about the causes of people's behavior.	**R**
P9	I have found that the causes of human behavior are usually complex rather than simple.	
P10	I am very interested in understanding how my own thinking works when I make judgments about people or attribute causes to their behavior.	
P11	I think very little about the different ways in which people influence one another.	**R**
P12	To understand a person's personality or behavior, I have found it important to know how their attitudes, beliefs, and character traits fit together.	
P13	When I try to explain another person's behavior, I focus on the person and do not worry too much about all the external factors that might be influencing them.	**R**
P14	I have often found that the basic cause of a person's behavior lies in the distant past.	
P15	I really enjoy analyzing the reasons or causes for people's behavior.	
P16	I generally think that complicated explanations for human behavior are more confusing than helpful.	**R**

#	Item	R?
P17	I reflect very little on how my thinking works in the process of understanding or explaining people's behavior.	**R**
P18	I think very little about the influence that other people have on my behavior.	**R**
P19	I have reflected quite a bit on how different parts of my personality influence one another (for example, beliefs affecting attitudes or attitudes affecting character traits).	
P20	I think a lot about the influence that society has on other people.	
P21	When I analyze a person's behavior, I often find that the causes form a chain going back through time, sometimes for years.	
P22	I am not very curious about human behavior.	**R**
P23	I prefer simple rather than complex explanations for people's behavior.	**R**
P24	When the reasons I give for my own behavior differ from another person's reasons, this makes me reflect on the thought processes that lead to my explanations.	
P25	I believe that to understand a person it is also necessary to understand the people with whom that person has close contact.	
P26	I tend to take people's behavior at face value and not worry too much about internal causes (e.g., attitudes, beliefs, traits, etc.).	**R**
P27	I think a lot about the influence that society has on my behavior and personality.	
P28	I have reflected very little on my own family background and personal history in order to understand why I am the kind of person I am.	**R**
–	*Attention:* "To assess your level of attention, please mark 'Agree' on this question."	–

Inclusion of Other in the Self (IOS) Perceived interpersonal closeness is measured using the IOS pictorial item (Aron et al., 1992). Participants choose one of seven pairs of increasingly overlapping circles labeled "Self" and "Other" to represent their closeness to a target (e.g., family, friends, strangers, classmates, or other social groups). The response is coded 1–7, with higher values indicating greater perceived closeness.

Each target is answered using a pictorial 7-option overlap scale between "Yo" and "Otros", like in the Figure A.4

Reading the Mind in the Eyes Test (RMET) Affective theory of mind is measured using the revised RMET (Baron-Cohen et al., 2001). Each item presents the eye region of a

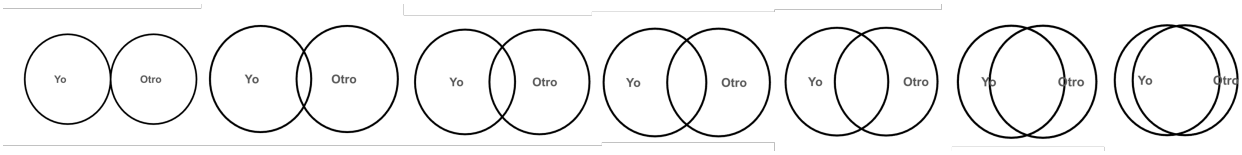


Figure A.4: Inclusion of the Other in the Self pictorial options

Table A.8: IOS targets shown in the survey

#	**Target (as shown)**
P1	Family
P2	Close friends
P3	Strangers (people you do not know)
P4	People with political views different from mine
P5	People from a higher socioeconomic stratum than mine
P6	People from a lower socioeconomic stratum than mine
P7	People of a different nationality than mine
P8	Neighbors
P9	Classmates

face and four candidate mental-state terms; participants select the term that best matches the expression. The RMET score is the number of correct responses, so higher values indicate higher performance.

A.2 A simple threshold social-weight model of altruism

A.2.1 Theoretical setup and behavioral implication

A decision-maker i chooses a prosocial action $a \in [0, \bar{a}]$ that benefits the other j and is costly to the self. Let own and other monetary payoffs be $\pi_i(a) = -c(a)$ and $\pi_j(a) = b(a)$. Assume $c'(a) > 0$, $c''(a) > 0$. Assume $b'(a) > 0$ and $b''(a) \leq 0$.

The decision-maker's utility is

$$u(a \mid \beta) \;=\; -c(a) \;+\; [\beta]_+ \, b(a), \qquad [\beta]_+ \equiv \max\{\beta, 0\}.$$

Here β is a latent "social weight" (other-regarding preference) drawn from representation social model s given by distribution F_s. If the draw is sufficiently high, the decision-maker weights the other payoff when deciding over action a.

For an interior solution (when $\beta > 0$), $a^*(\beta)$ satisfies

$$c'(a^*(\beta)) \;=\; \beta \, b'(a^*(\beta)). \tag{4}$$

Thus $a^*(\beta) = 0$ for $\beta \leq 0$, and $a^*(\beta)$ is increasing for $\beta > 0$.

Definition 1. *Let $\beta \sim F_C$ denote the draw under a coarse representation social model $s = C$ (control) and $\beta \sim F_T$ the draw under a fine representation social model $s = T$ (treatment). A fiction-reading treatment is a mean-preserving spread in other-regarding concerns:*

$$\mathbb{E}_T[\beta] = \mathbb{E}_C[\beta], \qquad F_T \text{ is a mean-preserving spread of } F_C.$$

This says fiction-reading treatment increases heterogeneity in other-regarding concern without changing its average level. In the simplest benchmark, control is a mass point $\beta \equiv \bar{\beta}$. Fiction exposure changes the *distribution* of β as the other becomes more psychologically textured, thus increasing the other-regarding preference weight. As (Brodsky, 1991) puts it "human diversity is literature's lock and stock (...) literature is the greatest—surely greater than any creed—teacher of human subtlety"

Proposition 1 (Mean-preserving spread in β raises expected prosociality)**.** *Assume $c'(a)$ is convex and b is concave. Then the policy function $\beta \mapsto a^*(\beta)$ is convex on $\mathbb{R}$. Consequently, if F_T is a mean-preserving spread of F_C as in Definition 1, then*

$$\mathbb{E}_T[a^*(\beta)] \;\geq\; \mathbb{E}_C[a^*(\beta)]\,,$$

with strict inequality whenever the spread is nondegenerate and the support of β intersects both $(-\infty, 0]$ and $(0, \infty)$.

Proof. By (4), $a^*(\beta)$ solves $c'(a) = [\beta]_+ b'(a)$. Under $c'''(a) \geq 0$ and concave b, the inverse mapping from $[\beta]_+$ to a^* is convex, and $[\beta]_+$ is convex in β which yields convexity of $a^*(\beta)$. A mean-preserving spread increases the expectation of any convex function, giving the result. □

Corollary 1 (Closed-form benchmark (quadratic costs))**.** *If $b(a) = a$ and $c(a) = \frac{1}{2}a^2$, then*

$$a^*(\beta) = [\beta]_+ \quad \textit{and} \quad \mathbb{E}[a^*(\beta)] = \mathbb{E}[[\beta]_+].$$

Therefore any mean-preserving spread in β (holding $\mathbb{E}[\beta]$ fixed) weakly increases expected prosociality, with strict increase whenever the spread is nondegenerate and the support of β intersects both $(-\infty, 0]$ and $(0, \infty)$.

A.2.2 Individuation and attribution complexity

The choice model pins down behavior through the latent parameter β. We now interpret β as a source of individuation or attribution-complexity on others and add a measurement layer, without changing the choice problem.

Let R be a categorical "reason" stated by the decision-maker when explaining or predicting another person's worth or behavior. We treat R as a coarse verbalization of the latent other-regarding preference.

Definition 2 (Reason reporting as a partition of β)**.** *There exists a finite set of reasons* $\mathcal{R} = \{0, 1, \ldots, M\}$ *and thresholds*

$$-\infty = \kappa_{-1} < \kappa_0 = 0 < \kappa_1 < \cdots < \kappa_M < +\infty$$

such that

$$R = g(\beta) \quad \textit{where} \quad g(\beta) = m \iff \beta \in (\kappa_{m-1}, \kappa_m].$$

Category $R = 0$ *corresponds to* $\beta \leq 0 \Rightarrow$ *categorical (non-individuated) processing,*

and $R \geq 1$ *corresponds to* $\beta > 0 \Rightarrow$ *individuated processing,*

and larger β *corresponds to richer, more fine-grained attributions.*

More generally, for any $m \in \{1, \ldots, M\}$,

$$\mathbb{P}_s(R \geq m) = \mathbb{P}_s(\beta > \kappa_{m-1}) \tag{5}$$

for each state $s \in \{C, T\}$.

Reasons Support expansion An implication of fiction-reading being a mean preserving spread of β is that the individual will expand the *range* of attribution categories used when thinking about other's motivation. It only requires that treatment expands the realized support of β enough to cross additional thresholds of the reporting partition.

Proposition 2 (Support expansion of attribution categories)**.** *Under Definition 2, suppose there exists at least one threshold* κ_m *such that treatment assigns positive probability to a bin* $(\kappa_{m-1}, \kappa_m]$ *that has zero probability under control. Then*

$$\text{supp}_T(R) \supsetneq \text{supp}_C(R).$$

Proof. Let $B_m \equiv (\kappa_{m-1}, \kappa_m]$ for $m \in \{0, 1, \ldots, M\}$, so that $R = m$ if and only if $\beta \in B_m$. Hence, for any state $s \in \{C, T\}$,

$$\mathbb{P}_s(R = m) = \mathbb{P}_s(\beta \in B_m).$$

Therefore,

$$m \in \text{supp}_s(R) \quad \iff \quad \mathbb{P}_s(\beta \in B_m) > 0.$$

By assumption, there exists at least one bin $B_{\tilde{m}}$ such that

$$\mathbb{P}_T(\beta \in B_{\tilde{m}}) > 0 \qquad \text{while} \qquad \mathbb{P}_C(\beta \in B_{\tilde{m}}) = 0.$$

Hence $\tilde{m} \in \text{supp}_T(R)$ but $\tilde{m} \notin \text{supp}_C(R)$, implying

$$\text{supp}_T(R) \supsetneq \text{supp}_C(R).$$

□

Extensive-margin individuation. A mean-preserving spread in β raises expected prosociality through the convexity of $[\beta]_+$, but it does not by itself imply a higher probability of crossing the individuation threshold $\beta = 0$. To obtain a prediction about the extensive margin of individuation, we therefore add an assumption on how treatment reallocates probability mass locally around zero.

Definition 3 (Local rightward shift near the individuation threshold)**.** *In addition to Definition 1, suppose there exists $\varepsilon > 0$ such that*

$$F_T(x) \leq F_C(x) \qquad \text{for all } x \in [-\varepsilon, \varepsilon],$$

with strict inequality for some x in that interval.

Proposition 3 (Treatment increases the probability of individuated attributions)**.** *Under Definitions 2 and 3,*

$$\mathbb{P}_T(\beta > 0) \geq \mathbb{P}_C(\beta > 0),$$

with strict inequality if $F_T(0) < F_C(0)$. Therefore

$$\mathbb{P}_T(R \geq 1) \geq \mathbb{P}_C(R \geq 1) \qquad \text{and} \qquad \mathbb{P}_T(R = 0) \leq \mathbb{P}_C(R = 0). \tag{6}$$

More generally, for any threshold category $m \in \{1, \ldots, M\}$,

$$\mathbb{P}_T(R \geq m) \geq \mathbb{P}_C(R \geq m)$$

whenever $\mathbb{P}_T(\beta > \kappa_{m-1}) \geq \mathbb{P}_C(\beta > \kappa_{m-1})$.

Proof. By Definition 3, evaluating at $x = 0$ gives

$$F_T(0) \leq F_C(0).$$

Since $F_s(0) = \mathbb{P}_s(\beta \leq 0)$, it follows that

$$\mathbb{P}_T(\beta \leq 0) \leq \mathbb{P}_C(\beta \leq 0),$$

and therefore

$$\mathbb{P}_T(\beta > 0) \geq \mathbb{P}_C(\beta > 0).$$

By Definition 2, $R = 0$ if and only if $\beta \leq 0$, while $R \geq 1$ if and only if $\beta > 0$. Hence

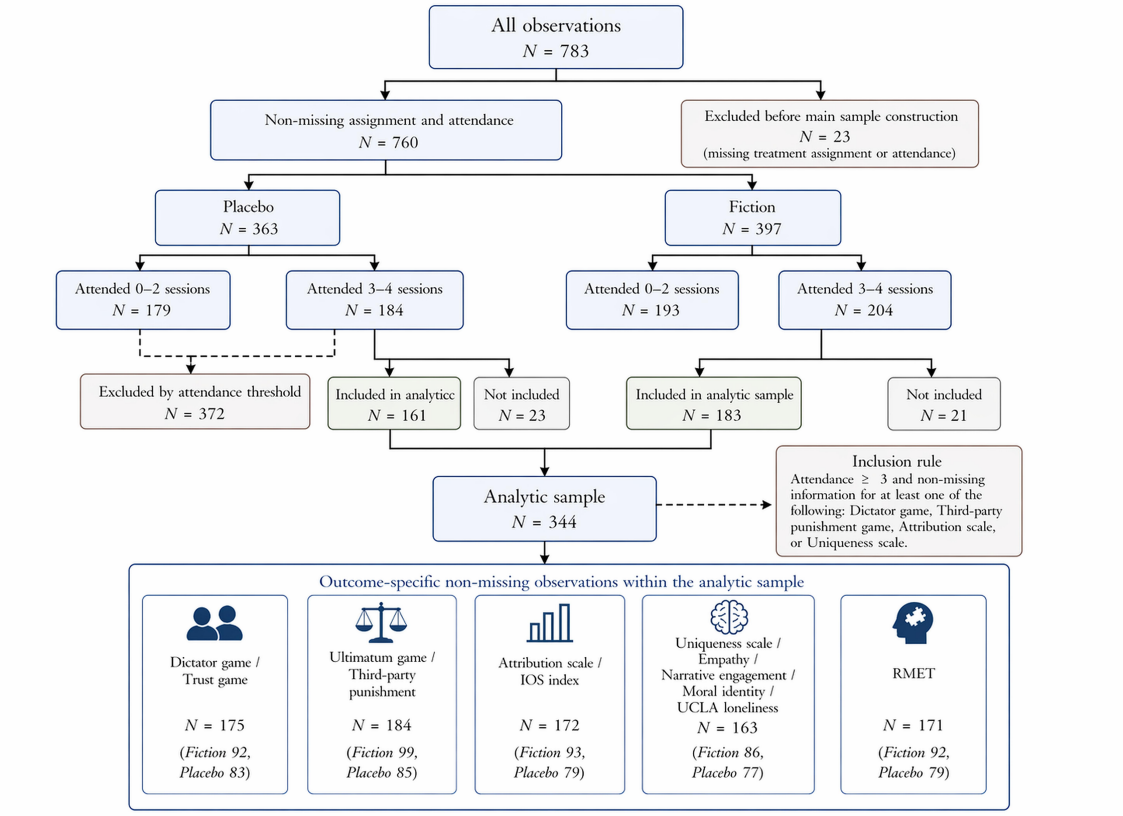
$$\mathbb{P}_T(R \geq 1) \geq \mathbb{P}_C(R \geq 1) \qquad \text{and} \qquad \mathbb{P}_T(R = 0) \leq \mathbb{P}_C(R = 0).$$

The final statement follows immediately from (5). □

A.3 Additional results and robustness checks

A.3.1 Sample construction and data availability

Figure A.5: Sample construction and outcome-specific availability



Notes: The figure summarizes the construction of the analytic sample and the resulting outcome-specific sample sizes. Counts are reported at the observation level. The analytic sample is defined by attendance of at least three course sessions and non-missing information for at least one of the inclusion-relevant measures: the Dictator game, the Third-party punishment game, the Attribution scale, or the Uniqueness scale. Outcome-specific sample sizes are smaller than the analytic sample because the inclusion rule is based on an OR condition: an observation may enter the analytic sample through one outcome family even if another specific outcome is missing.

Figure A.5 summarizes how the analytic sample is formed from treatment assignment, attendance, and outcome availability. The main source of sample loss is limited attendance: observations associated with participants who attended only 0–2 course sessions are excluded by the attendance threshold, while observations associated with 3–4 attended sessions are eligible for inclusion if they contain non-missing information for at least one inclusion-relevant measure. Table A.9 further clarifies why sufficient course attendance does not mechanically imply inclusion in the analytic sample. Among observations with at least three attended course sessions, 16.2% do not contribute information in any experimental measurement session and are therefore excluded from the analytic sample. The remaining observations

Table A.9: Experimental measurement participation among observations with sufficient course attendance

	Placebo		Fiction		Total	
Measurement participation	N	%	N	%	N	%
Panel A. Number of experimental measurement sessions						
0 sessions	23	18.1	21	14.6	44	16.2
1 session	47	37.0	63	43.8	110	40.6
2 sessions	57	44.9	60	41.7	117	43.2
Panel B. Timing of experimental measurement						
Short-term measurement	98	77.2	115	79.9	213	78.6
Long-term measurement	63	49.6	68	47.2	131	48.3

Notes: This table reports experimental measurement participation among observations with at least three attended course sessions. Panel A reports the number of experimental measurement sessions in which an observation contributes non-missing information to at least one inclusion-relevant measure. These categories are mutually exclusive. Panel B reports whether an observation contributes non-missing information in the short-term or long-term measurement horizon. These categories are not mutually exclusive, since an observation may contribute information in both horizons. Short term refers to the immediate post-course measurement session, while long term refers to the follow-up measurement session. Percentages are computed within treatment arm. Observations with zero experimental measurement sessions are excluded from the analytic sample.

contribute information in one measurement horizon (40.6%) or in both horizons (43.2%). Measurement participation is broadly similar across the fiction and placebo arms.

Table A.10: Additional baseline balance: reading habits and prior beliefs

	(1) Mean	(2) Sd	(3) Treated	(4) Placebo	(5) Pvalue
Reading hours per week (# of hours)	4.16	4.21	4.55	3.74	0.07
Fiction books read (# of books)	2.81	4.11	3.12	2.48	0.17
Non-fiction books read (# of books)	1.69	2.35	1.82	1.53	0.27
Work- or study-related books read (# of books)	5.31	8.09	5.39	5.22	0.85
Fiction books purchased (# of books)	1.77	3.50	1.65	1.90	0.54
Non-fiction books purchased (# of books)	1.00	2.00	1.11	0.89	0.34
Work- or study-related books purchased (# of books)	0.84	1.90	0.87	0.80	0.76
Fiction increases empathy (likert 1-4)	2.89	0.69	2.85	2.94	0.27
Fiction increases intelligence (likert 1-4)	2.74	0.58	2.73	2.74	0.84
Fiction helps understand complex motivations (likert 1-4)	3.03	0.72	3.04	3.01	0.73
Fiction helps recognize individual uniqueness (likert 1-4)	2.85	0.73	2.80	2.92	0.13
Observations			183	161	

Notes: This table reports additional baseline balance checks for variables not included in the main balance table. The variables capture pre-treatment reading habits, prior exposure to different types of books, and beliefs about the effects of fiction reading. Columns report the overall mean and standard deviation, the mean among treated and placebo participants, and the p-value for the difference in means across groups.

Table A.11: Player 1 Personal and Social Norms Regarding lack of trustworthiness

	Personal norm		Normative expectations	
	(1)	(2)	(3)	(4)
Treatment	-0.058	-0.079	0.001	-0.037
	(0.120)	(0.135)	(0.059)	(0.028)
Constant	2.494***	3.286***	2.314***	2.509**
	(0.108)	(0.397)	(0.031)	(0.889)
Placebo mean	2.494	2.494	2.313	2.313
Observations	175	171	175	171
R^2	0.001	0.073	0.000	0.048
Wave fixed effects	✓	✓	✓	✓
Constants		✓		✓

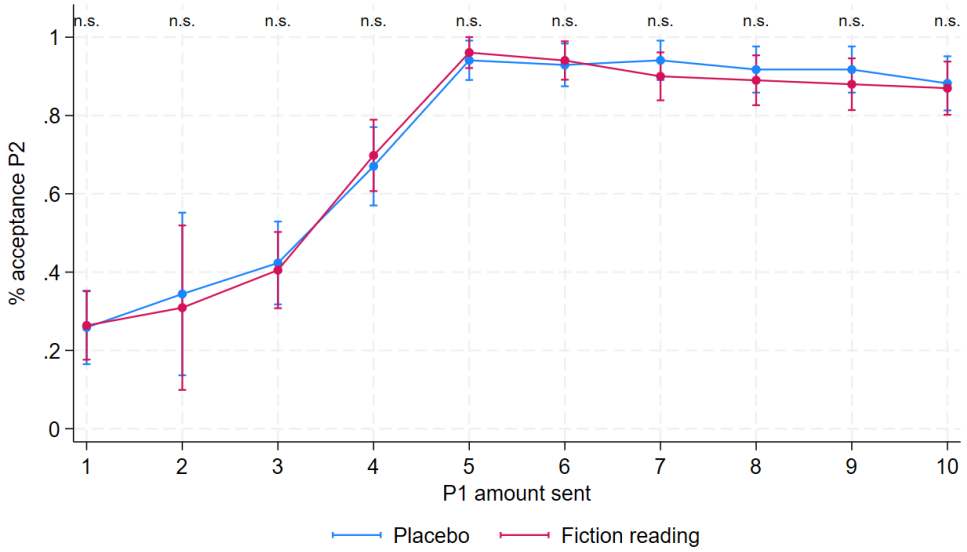
Notes: Entries report coefficients from linear regressions of trust-game norm- and expectation-related outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are (i) participants' personal norm regarding how much Player 2 should return in the trust game, and (ii) participants' empirical expectations about how much Player 2 would return conditional on the amount sent by Player 1. Personal norms are measured on a 1–4 scale ranging from "very inappropriate" to "very appropriate," capturing respondents' own evaluation of Player 2's return decision. Expectations are elicited on a 1–10 scale, indicating how much respondents believe Player 2 would return given the amount previously sent by Player 1. In the trust game, Player 1 decided how many tokens (0–10) to send to Player 2, and the amount sent was tripled before reaching Player 2. The questions in this table therefore refer to Player 2's return decision conditional on the transfer received from Player 1. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.12: Player 1 (Ultimatum) Behavior and Empirical Expectations

	Amount sent (P1)		P2 expectations about P1 sent	
	(1)	(2)	(3)	(4)
Treatment	0.262*	0.207	0.061	0.006
	(0.137)	(0.145)	(0.135)	(0.154)
Constant	4.652***	4.225***	4.353***	3.998***
	(0.097)	(0.876)	(0.084)	(0.876)
Placebo mean	4.647	4.647	4.353	4.353
Observations	184	180	184	180
R^2	0.015	0.072	0.001	0.056
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of ultimatum-game outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are: (i) the offer made by Player 1 (proposer), measured as the number of tokens (0–10) offered out of a 10-token endowment, and (ii) participants' incentivized empirical belief about the typical offer made by proposers in the ultimatum game. Beliefs were incentivized: participants received a monetary payment if their guess matched the modal response in the corresponding belief elicitation task. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.6: Responder Acceptance Rate in the Ultimatum Game by treatment status.



Notes: The figure plots the mean acceptance rate of Player 2 (responder) in the ultimatum game, separately by treatment status and by the amount offered by Player 1 (proposer). The dependent variable is an indicator equal to one if Player 2 accepts the offer X made by Player 1, elicited via the strategy method, and zero otherwise. In the ultimatum game, Player 1 proposes a transfer (0–10 tokens) to Player 2, and Player 2 decides whether to accept or reject; if Player 2 rejects, both players receive zero. Points correspond to fitted values from a fully saturated model interacting treatment assignment with the amount offered by Player 1 and including fixed effects for the first experimental wave. Vertical bars denote 95% confidence intervals. Standard errors clustered at the treatment session term level are reported in the confidence intervals. Significance markers above each offer indicate tests of equality between the fiction-reading treatment and placebo at that specific offer level: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, and n.s. indicates not statistically significant at the 10% level.

Table A.13: Player 2 (Ultimatum) Acceptance of Low Offers and Empirical Expectations

	P2 acceptance rate for offers < 5		P1 expectations about P2 accepting < 5	
	(1)	(2)	(3)	(4)
Treatment	0.015	0.021	-0.025	-0.020
	(0.040)	(0.039)	(0.059)	(0.052)
Constant	0.347***	0.382**	0.432***	0.197
	(0.024)	(0.169)	(0.041)	(0.344)
Control mean	0.349	0.349	0.435	0.435
Observations	184	180	184	180
R^2	0.016	0.056	0.016	0.046
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of ultimatum-game outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are: (i) Player 2's acceptance rate for offers smaller than 5 tokens, measured as the share of low offers accepted among observed responses for offers of 0, 1, 2, 3, and 4 tokens, and (ii) Player 1's expectation about whether Player 2 would accept an offer smaller than 5 tokens. In the ultimatum game, Player 1 (proposer) decided how many tokens to offer to Player 2 (responder) out of a 10-token endowment. The variable *P2 acceptance rate for offers* < 5 captures the responder's willingness to accept low offers. The variable *P1 expectations about P2 accepting* < 5 captures the proposer's empirical belief about responder behavior in low-offer situations. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.14: Responder Acceptance Decisions

	(1)	(2)
Treatment	0.022	0.035
	(0.051)	(0.052)
P1 Offer	0.087***	0.089***
	(0.006)	(0.006)
Treatment × P1 Offer	-0.005	-0.008
	(0.008)	(0.008)
Constant	0.229***	0.222***
	(0.038)	(0.037)
Placebo mean	0.686	0.686
Observations	1877	1836
R^2	0.33	0.34
Wave fixed effects	✓	✓
Controls		✓

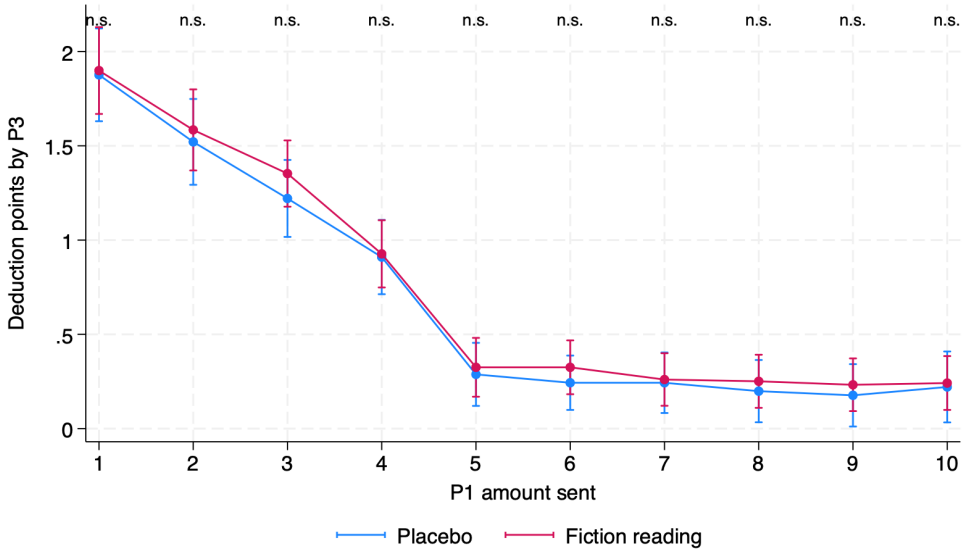
Notes: Entries report coefficients from linear regressions of ultimatum-game acceptance outcomes on an indicator for assignment to the fiction-reading treatment, the offer made by Player 1, and their interaction. The dependent variable is an indicator equal to one if Player 2 accepts the offer and zero otherwise. In the ultimatum game, Player 1 (proposer) decided how many tokens (0–10) to offer to Player 2 (responder) out of a 10-token endowment. The variable *P1 Offer* therefore measures the size of the proposed transfer to Player 2. The interaction term captures whether the relationship between the offer size and the probability of acceptance differs across treatment and control groups. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.15: Player 1 (Third party) Behavior and Empirical Expectations

	Amount sent		Expectations	
	(1)	(2)	(3)	(4)
Treatment	0.035	0.026	-0.216	-0.203
	(0.113)	(0.146)	(0.154)	(0.202)
Constant	4.595***	4.453***	4.464***	3.059***
	(0.078)	(0.744)	(0.061)	(0.714)
Placebo mean	4.600	4.600	4.471	4.471
Observations	184	180	184	180
R^2	0.004	0.026	0.011	0.033
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of third-party allocation-game outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are (i) the amount sent by Player 1 and (ii) Player 2's empirical expectations about the typical amount sent by Player 1. In the third-party allocation game, Player 1 decided how many tokens (0–10) to send to the relevant counterpart under the experimental allocation setting. The variable *Amount sent* therefore measures the observed transfer chosen by Player 1, while *Expectations* captures respondents' beliefs about the transfer that Player 1 would typically choose in the same setting. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figure A.7: Deduction Points Assigned by Player 3 to Player 1 by Treatment Status



Notes: The figure plots the mean number of deduction points assigned by Player 3 (third party) in the third-party punishment game, separately by treatment status and by the amount sent by Player 1. The dependent variable is the number of points deducted by Player 3 from Player 2's payoff. Points correspond to fitted values from a fully saturated model interacting treatment assignment with the amount sent by Player 1 and including fixed effects for the first experimental wave. Vertical bars denote 95% confidence intervals. Standard errors clustered at the treatment session term level are reported in the confidence intervals. Significance markers above each amount indicate tests of equality between the fiction-reading treatment and placebo at that specific amount sent: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, and n.s. indicates not statistically significant at the 10% level.

Table A.16: Deduction based of amount sent to others

	Deduction chosen by P3		Expected deduction by P2	
	(1)	(2)	(3)	(4)
	Baseline	With controls	Baseline	With controls
Treatment	0.082	0.022	0.019	-0.042
	(0.191)	(0.191)	(0.183)	(0.185)
Amount sent (P1)	-0.198***	-0.201***	-0.208***	-0.210***
	(0.019)	(0.019)	(0.020)	(0.020)
Treatment × Amount sent (P1)	-0.003	-0.001	-0.001	0.001
	(0.026)	(0.027)	(0.026)	(0.027)
Placebo mean	0.753	0.762	0.832	0.842
Observations	2021	1977	2024	1980
R^2	0.292	0.314	0.314	0.332
Wave fixed effects	✓	✓	✓	✓
Demographic controls		✓		✓

Notes: Entries report coefficients from linear regressions of third-party punishment outcomes on an indicator for assignment to the fiction-reading treatment, the amount sent by Player 1, and their interaction. The dependent variable in Columns (1)–(2) is the deduction chosen by Player 3. The dependent variable in Columns (3)–(4) is the deduction that respondents expect Player 3 would impose. In all columns, *Amount sent (P1)* refers to the observed amount sent by Player 1 in the underlying allocation stage. In the third-party punishment game, Player 1 makes an allocation decision, Player 2 is the affected party, and Player 3 may reduce Player 1's payoff in response to that allocation. The interaction term captures whether the relationship between the amount sent by Player 1 and the punishment imposed (or expected) by Player 3 differs across treatment and control groups. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.17: Player 2 (Third party) Personal and social Norms regarding Player 1

	Personal norm		Social norm	
	(1)	(2)	(3)	(4)
Treatment	-0.032	-0.011	-0.120	-0.096
	(0.071)	(0.070)	(0.109)	(0.118)
Constant	2.001***	1.927***	1.847***	1.489***
	(0.049)	(0.350)	(0.104)	(0.460)
Placebo mean	1.976	1.976	1.824	1.824
Observations	184	180	184	180
R^2	0.272	0.310	0.282	0.328
Wave fixed effects	✓	✓	✓	✓
Demographic controls		✓		✓

Notes: Entries report coefficients from linear regressions of norm-related outcomes in the third-party punishment game on an indicator for assignment to the fiction-reading treatment. The dependent variable in Columns (1)–(2) is *Personal norm*, defined as the respondent's own evaluation of the appropriateness of the punishment decision in the allocation scenario. The dependent variable in Columns (3)–(4) is *Social norm*, defined as the respondent's belief about how other participants evaluate the same punishment decision. Personal norms are measured on a 1–4 scale ranging from "very inappropriate" to "very appropriate." Social norms are elicited using the coordination-based method of Krupka and Weber (2013), in which participants receive a monetary bonus if their response matches the modal answer among other participants. In the third-party punishment game, Player 1 makes an allocation decision and Player 3 can reduce Player 1's payoff in response to that choice. The norm questions therefore capture respondents' evaluations of punishment in that third-party allocation setting. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.18: Player 2 (Third party) Personal and social Norms regarding Player 3

	Pnorm deductions of P3		Snorm deductions of P3	
	(1)	(2)	(3)	(4)
Treatment	-0.227*	-0.262**	-0.094	-0.085
	(0.107)	(0.117)	(0.088)	(0.086)
Constant	3.030***	4.002***	2.974***	4.059***
	(0.076)	(0.286)	(0.072)	(0.368)
Placebo mean	3.000	3.000	2.953	2.953
Observations	184	180	184	180
R^2	0.315	0.372	0.181	0.214
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of norm-related outcomes in the third-party punishment game on an indicator for assignment to the fiction-reading treatment. The dependent variable in Columns (1)–(2) is the participant's *personal norm* regarding the deduction chosen by Player 3, that is, the respondent's own evaluation of the appropriateness of Player 3's punishment decision. The dependent variable in Columns (3)–(4) is the *social norm* regarding deductions by Player 3, that is, the respondent's belief about how other participants evaluate the same punishment decision. Personal norms are measured on a 1–4 scale ranging from "very inappropriate" to "very appropriate." Social norms are elicited using the coordination-based method of Krupka and Weber (2013), in which participants receive a monetary bonus if their response matches the modal answer among other participants. In the third-party punishment game, Player 1 makes an allocation decision, Player 2 is the affected party, and Player 3 may reduce Player 1's payoff in response to that allocation. The norm questions in this table therefore refer specifically to the deduction imposed by Player 3. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.19: Social connectedness

	(1)	(2)
Treatment	0.044	0.037
	(0.065)	(0.072)
Constant	2.345***	3.033***
	(0.051)	(0.320)
Placebo mean	2.344	2.344
Observations	163	160
R^2	0.004	0.075
Wave fixed effects	✓	✓
Constants		✓

Notes: Entries report coefficients from linear regressions of perceived social isolation on an indicator for assignment to the fiction-reading treatment. The dependent variable is constructed as the mean of items from the UCLA Loneliness Scale (Version 3) (Russell, 1996). Following the original coding procedure, positively worded items (Items 5, 6, 9, 10, 15, 16, 19, and 20) are reverse-coded as $5 - x$, where responses are measured on a 1–4 scale, so that higher values consistently indicate greater perceived loneliness. The scale captures subjective feelings of social isolation and lack of connectedness. See Appendix Table A.4 for the full list of items and coding details. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.3.2 Other channels: Moral self-concept

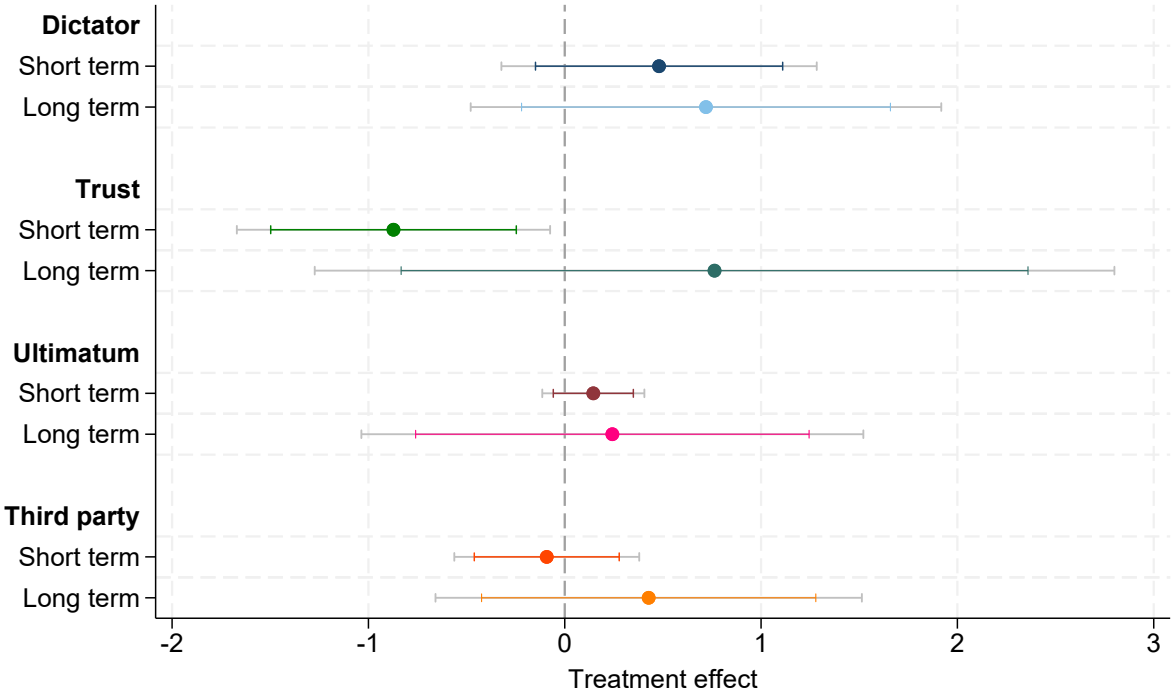
Table A.20: Morality indices

	Symbolization		Internalization	
	(1)	(2)	(3)	(4)
Treatment	0.066	0.171	-0.011	-0.002
	(0.171)	(0.140)	(0.059)	(0.069)
Constant	4.193***	3.265***	5.422***	4.973***
	(0.155)	(0.772)	(0.039)	(0.657)
Placebo mean	4.188	4.188	5.421	5.421
Observations	163	160	163	160
R^2	0.012	0.109	0.001	0.073
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of Moral Identity measures on an indicator for assignment to the fiction-reading treatment. The dependent variables correspond to the two standard dimensions of the Moral Identity Scale (Aquino & Reed, 2002): *Internalization*, which captures the extent to which moral traits are central to the self-concept, and *Symbolization*, which captures the degree to which individuals express their moral identity publicly. The *Symbolization* index is defined as the mean of items 10–13, which measure the degree to which moral traits are reflected in a person's public self-presentation. It captures whether being a moral person is expressed through actions and symbols that are observable to others. The *Internalization* index is defined as the mean of items 1–6, which measure the degree to which being a moral person is an important and deeply held part of an individual's identity, regardless of whether it is visible to others. See Appendix Table A.6 for the full list of items and coding details. All specifications include fixed effects for the first experimental wave. Even-numbered columns additionally include demographic controls: age, quiz score, an indicator for short-term sessions, prior acquaintance with other participants, gender, an indicator for high socioeconomic stratum, and an indicator for economics as field of study. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

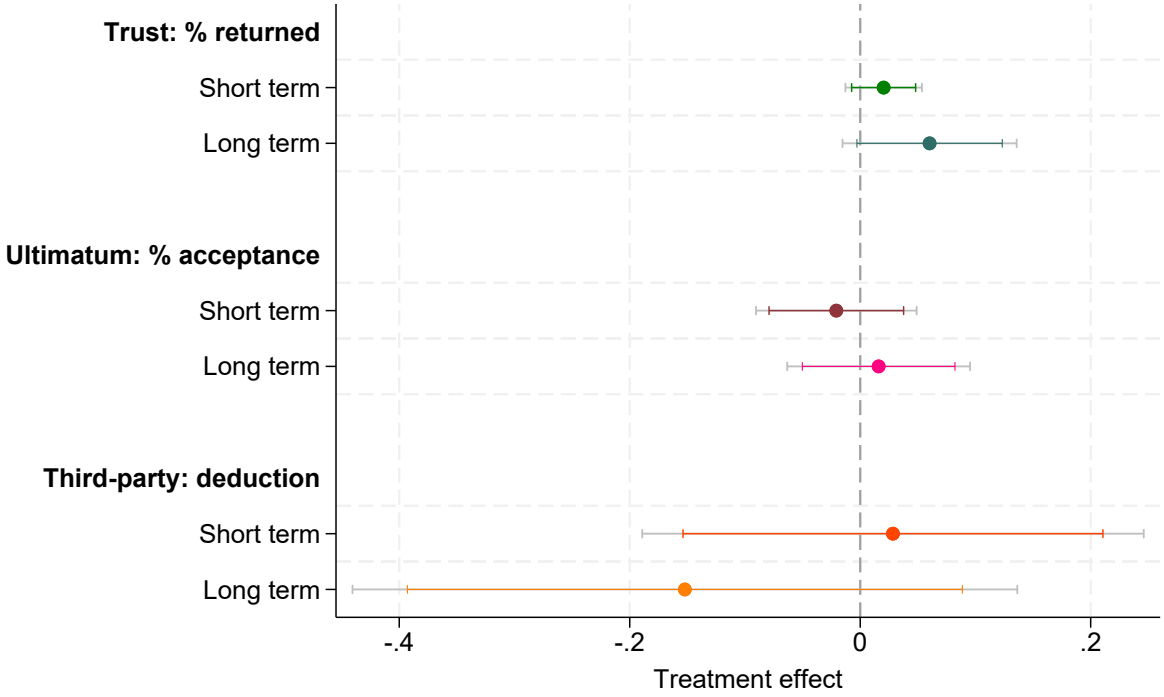
A.3.3 Short vs. Long term treatment effects

Figure A.8: Treatment effects of behavior in games (Short term vs Long term)



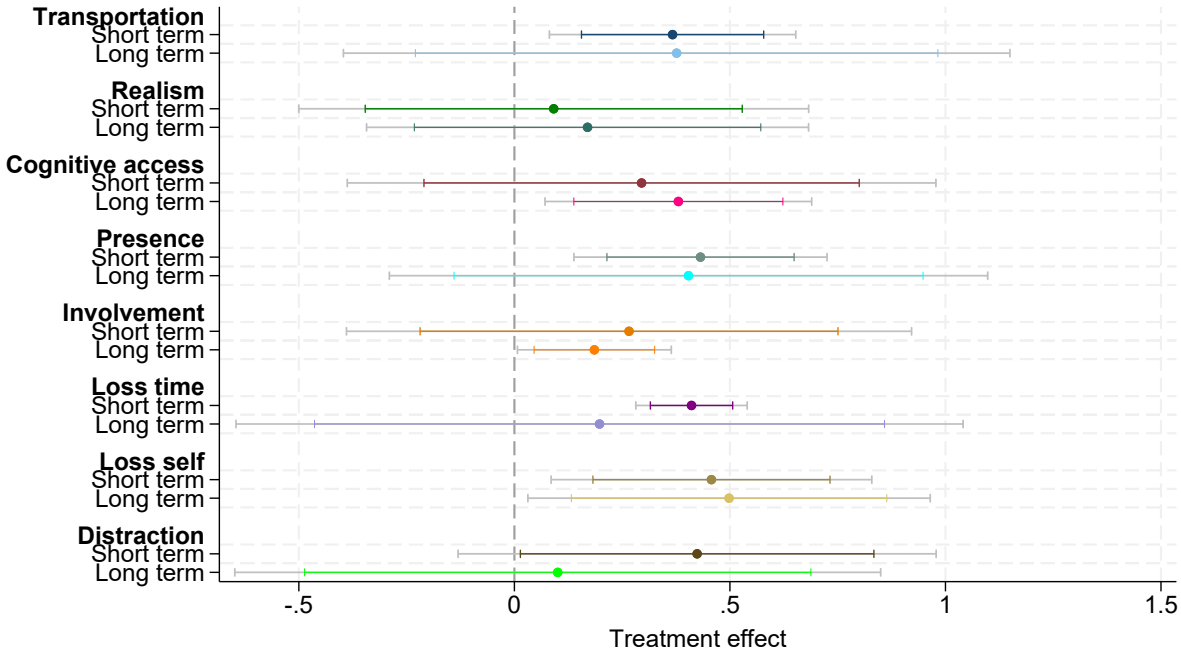
Notes: The figure reports treatment-effect estimates from separate linear regressions of behavioral outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for participants surveyed in the short term and in the long term after treatment assignment. All specifications include first-wave fixed effects and standard errors are clustered at the session level. Markers represent point estimates colored horizontal bars denote 90% and grey horizontal bars denote 95% confidence intervals. Behavioral outcomes correspond to the main decision variable in each game.

Figure A.9: Treatment effects of behavior in games (Short term vs Long term)



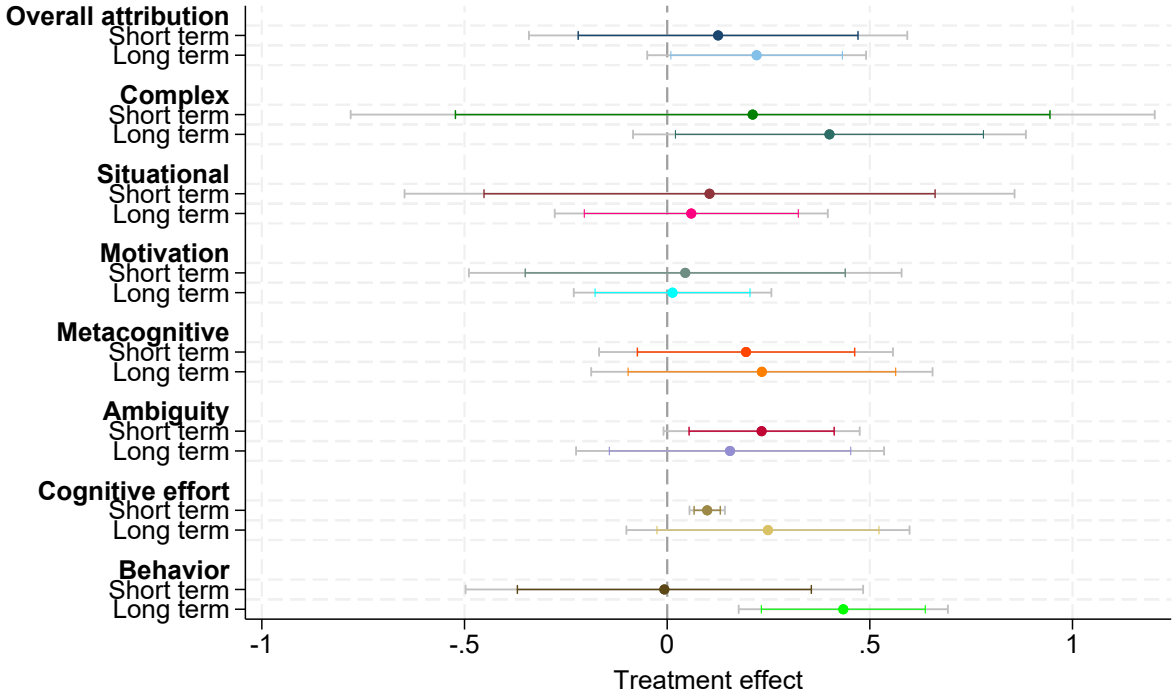
Notes: The figure reports treatment-effect estimates from separate linear regressions of panel outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. In the trust game, the outcome is the share of the received amount returned by Player 2, constructed from conditional return choices for each amount sent by Player 1. In the ultimatum game, the outcome is an indicator for accepting a given offer, estimated across the full set of possible offers. In the third-party punishment game, the outcome is the deduction chosen for each value received by Player 3. All specifications include first-wave fixed effects and fixed effects for the relevant conditional-choice level (amount sent, offer, or amount received). Standard errors are clustered at the respondent-by-horizon level (`id_obs_term`). Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.10: Treatment effects on Narrative subscales (Short term vs Long term)



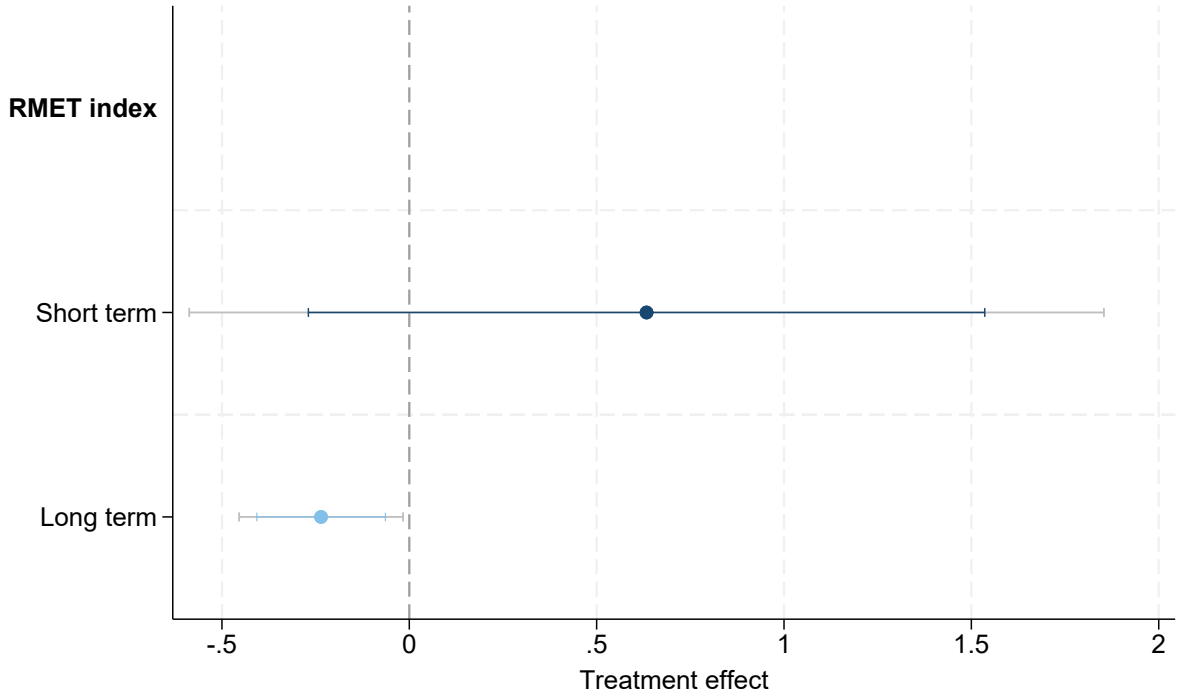
Notes: The figure reports treatment-effect estimates from separate baseline regressions of narrative-scale outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.11: Treatment effects on Attribution indices (Short term vs Long term)



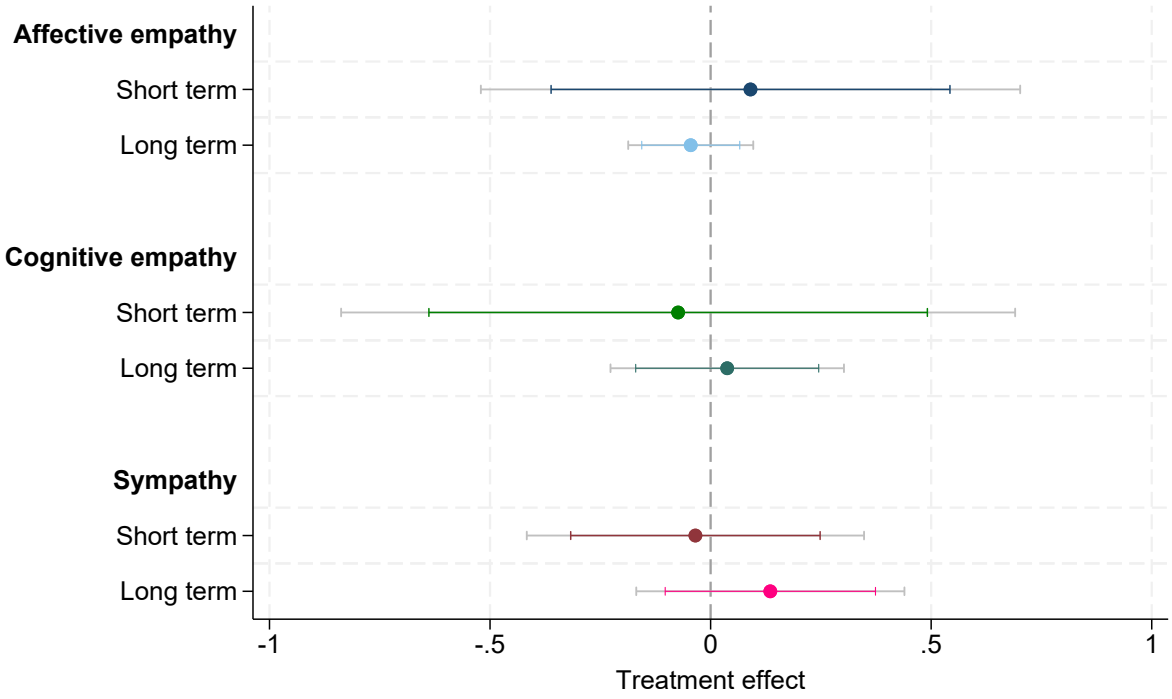
Notes: The figure reports treatment-effect estimates from separate baseline regressions of attribution outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. Outcomes include the overall attribution index and the seven attribution subscales: Complex, Situational, Motivation, Metacognitive, Ambiguity, Cognitive effort, and Behavior. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.12: Treatment effects on Theory of Mind (Short term vs Long term)



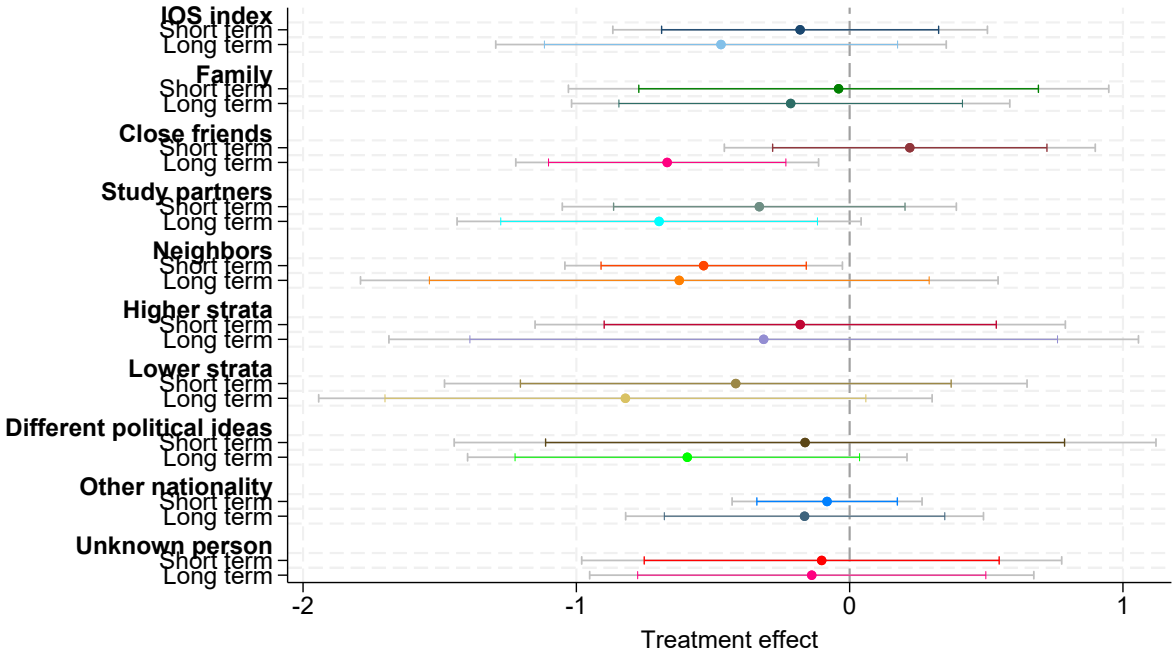
Notes: The figure reports treatment-effect estimates from separate baseline regressions of RMET accuracy on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.13: Treatment effects on Empathy indices (Short term vs Long term)



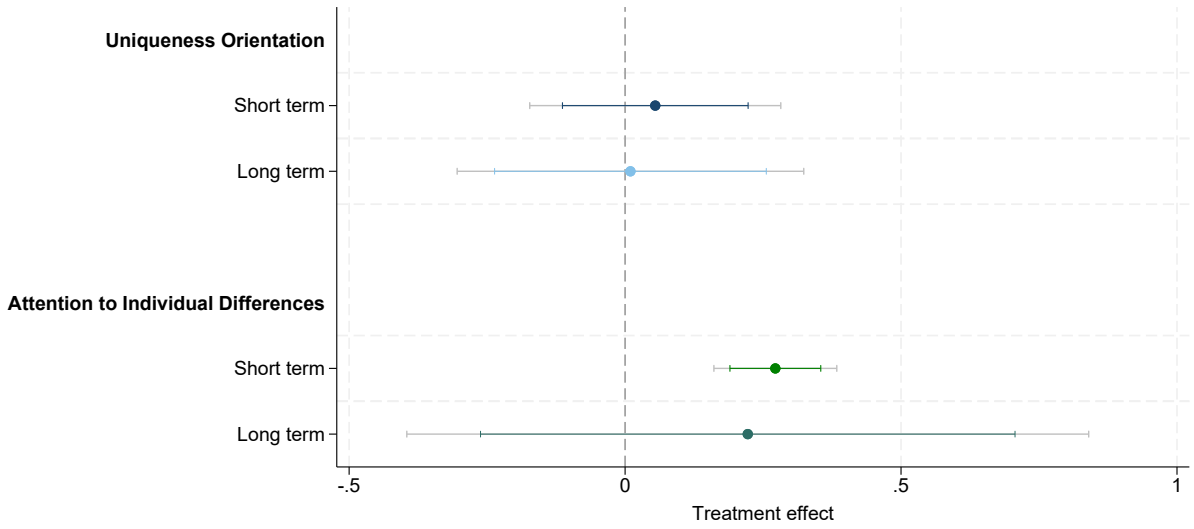
Notes: The figure reports treatment-effect estimates from separate baseline regressions of empathy outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.14: Treatment effects on IOS indices (Short term vs Long term)



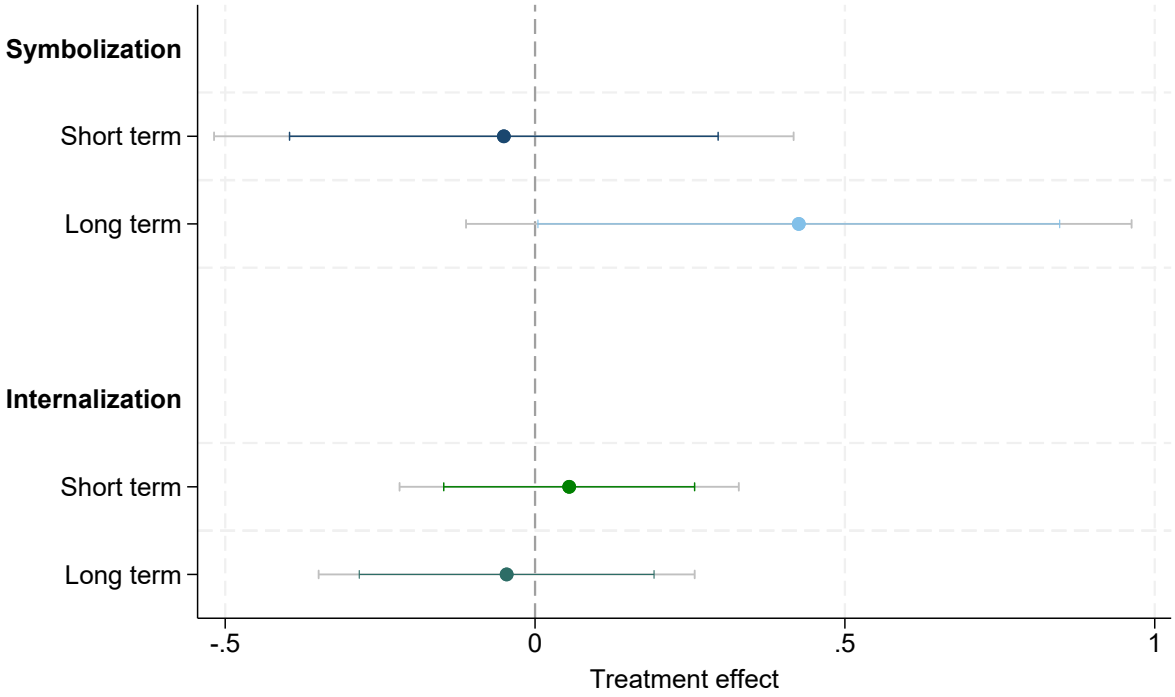
Notes: The figure reports treatment-effect estimates from separate baseline regressions of IOS outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. Outcomes include the overall IOS index and its component items. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.15: Treatment effects on Uniqueness (Short term vs Long term)



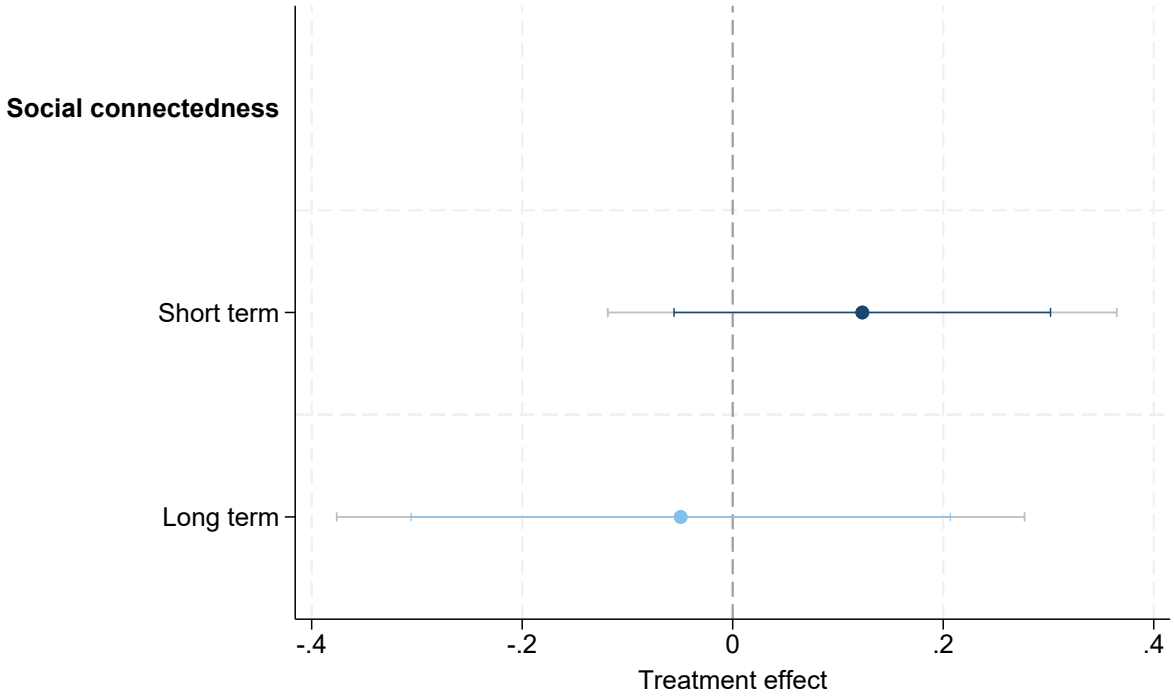
Notes: The figure reports treatment-effect estimates from separate baseline regressions of uniqueness outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.16: Treatment effects on Morality (Short term vs Long term)



Notes: The figure reports treatment-effect estimates from separate baseline regressions of morality outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. Outcomes correspond to the Symbolization (`idx_moral_symb`) and Internalization (`idx_moral_int`) subscales. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

Figure A.17: Treatment effects on Social connectedness (Short term vs Long term)



Notes: The figure reports treatment-effect estimates from separate baseline regressions of the UCLA index on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and long-term samples. All specifications include first-wave fixed effects, and standard errors are clustered at the session-by-horizon level. Markers represent point estimates. Gray horizontal bars denote 95% confidence intervals, and colored horizontal bars denote 90% confidence intervals.

A.3.4 Randomization inference and Multiple hypothesis testing

In this appendix, we report two complementary robustness exercises for the main experimental outcomes: randomization inference and multiple-hypothesis adjustment. Randomization inference (RI) re-estimates the treatment effect under repeated reassignments of treatment status consistent with the experimental design, thereby providing p-values that rely directly on the randomization procedure rather than only on large-sample approximations. Multiple-hypothesis testing (MHT) addresses the fact that the paper evaluates treatment effects across several related outcomes, which increases the likelihood of false positives when each hypothesis is considered in isolation.

For the main robustness tables, we compute Benjamini–Hochberg adjusted p-values within conceptual families that correspond to the main claims of the paper. The first family, *core prosocial behavior*, includes dictator-game giving and trust-game trustworthiness. The second family, *secondary behavioral outcomes*, includes trust-game sending, ultimatum-game offers, and third-party punishment behavior. The third family, *norm-related outcomes*, includes empirical expectations, personal norms, and social norms across the behavioral games. The fourth family, *focal mechanisms*, includes outcomes related to individuation and richer social representation: uniqueness orientation, attention to individual differences, and attribution complexity. The fifth family, *other psychological and rival channels*, includes transportation, RMET, affective empathy, cognitive empathy, moral identity, UCLA loneliness, and the IOS summary measure. This structure aligns the multiplicity correction with the substantive claims tested in the paper while avoiding overly broad corrections across conceptually unrelated outcomes.

We also report disaggregated robustness tables for selected scale batteries and construct-specific measures. These tables adjust p-values within each scale battery or construct, including narrative engagement, empathy, uniqueness, attribution, moral identity, IOS, UCLA, and RMET. They are intended to provide transparency on the components of each construct rather than to define the main inferential families of the paper.

In the tables below, each outcome reports three p-values: the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value in braces, and the randomization-inference p-value in brackets.

Table A.21: RI and MHT on family 1

	Dictator giving		Trustworthiness	
	(1)	(2)	(3)	(4)
Treatment	(0.030)	(0.049)	(0.119)	(0.140)
	{0.059}	{0.099}	{0.119}	{0.140}
	[0.053]	[0.101]	[0.111]	[0.243]
Observations	175	171	138	135
First-wave FE	✓	✓	✓	✓
Controls		✓		✓

Notes: This table reports robustness results for the core prosocial behavior family. The two column groups correspond to: (i) *Dictator giving*, measured as the amount sent by Player 1 in the dictator game; and (ii) *Trustworthiness*, measured as the average share returned by Player 2 across conditional return decisions in the trust game. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the core prosocial behavior family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.22: RI and MHT on family 2

	Trust sending		Ultimatum offer		Third-party punishment	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	(0.458)	(0.326)	(0.081)	(0.180)	(0.760)	(0.860)
	{0.688}	{0.489}	{0.244}	{0.489}	{0.760}	{0.860}
	[0.559]	[0.494]	[0.134]	[0.341]	[0.795]	[0.918]
Observations	175	171	184	180	184	180
First-wave FE	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓

Notes: This table reports robustness results for the secondary behavioral outcomes family. The three column groups correspond to: (i) *Trust sending*, measured as the amount sent by Player 1 in the trust game; (ii) *Ultimatum offer*, measured as the amount offered by Player 1 in the ultimatum game; and (iii) *Third-party punishment*, measured as the punishment chosen in the third-party punishment game. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the secondary behavioral outcomes family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.23: RI and MHT on family 3

	DG expectations		DG personal norm		DG social norm		TG expectations		TG J2 personal norm		TG J2 social norm		TG J1 personal norm		TG J1 social norm		UG expectations		TPP expectations		TPP J2 personal norm		TPP J2 social norm		TPP J3 personal norm		TPP J3 social norm	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)	(25)	(26)	(27)	(28)
Treatment	(0.259)	(0.203)	(0.654)	(0.806)	(0.907)	(0.845)	(0.810)	(0.949)	(0.022)	(0.009)	(0.029)	(0.014)	(0.638)	(0.570)	(0.983)	(0.212)	(0.660)	(0.971)	(0.189)	(0.335)	(0.657)	(0.879)	(0.294)	(0.430)	(0.057)	(0.047)	(0.308)	(0.344)
	{0.615}	{0.594}	{0.840}	{0.971}	{0.977}	{0.971}	{0.946}	{0.971}	{0.204}	{0.096}	{0.204}	{0.096}	{0.840}	{0.887}	{0.983}	{0.594}	{0.840}	{0.971}	{0.615}	{0.688}	{0.840}	{0.971}	{0.615}	{0.752}	{0.267}	{0.218}	{0.615}	{0.688}
	[0.352]	[0.350]	[0.719]	[0.853]	[0.904]	[0.853]	[0.848]	[0.962]	[0.100]	[0.138]	[0.087]	[0.135]	[0.709]	[0.633]	[0.989]	[0.374]	[0.758]	[0.981]	[0.333]	[0.489]	[0.717]	[0.917]	[0.367]	[0.477]	[0.172]	[0.234]	[0.416]	[0.531]
Observations	175	171	175	171	175	171	175	171	175	171	175	171	175	171	175	171	184	180	184	180	184	180	184	180	184	180	184	180
First-wave FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓

Notes: This table reports robustness results for the norm-related outcomes family. The fourteen column groups correspond to empirical expectations, personal norms, and social norms measured across the dictator game, trust game, ultimatum game, and third-party punishment game. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the norm-related outcomes family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.24: RI and MHT on family 4

	Uniqueness		Attention to individual differences		Attribution complexity	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	(0.616)	(0.547)	(0.048)	(0.031)	(0.143)	(0.130)
	{0.616}	{0.547}	{0.143}	{0.094}	{0.214}	{0.196}
	[0.660]	[0.642]	[0.145]	[0.171]	[0.217]	[0.269]
Observations	163	160	163	160	172	169
First-wave FE	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓

Notes: This table reports robustness results for the focal mechanism family. The three column groups correspond to: (i) *Uniqueness*, measuring orientation toward individual distinctiveness; (ii) *Attention to individual differences*; and (iii) *Attribution complexity*, measuring the tendency to represent others' behavior in more complex and less one-dimensional terms. These outcomes are grouped together because they capture the paper's proposed mechanism of individuation and richer social representation. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the focal mechanism family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.25: RI and MHT on Family 5

	Transportation		RMET		Affective empathy		Cognitive empathy		Moral symbolization		Moral internalization		UCLA loneliness		IOS summary	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Treatment	(0.064)	(0.045)	(0.442)	(0.357)	(0.893)	(0.764)	(0.663)	(0.890)	(0.707)	(0.248)	(0.861)	(0.978)	(0.513)	(0.615)	(0.065)	(0.072)
	{0.258}	{0.287}	{0.893}	{0.713}	{0.893}	{0.978}	{0.893}	{0.978}	{0.893}	{0.662}	{0.893}	{0.978}	{0.893}	{0.978}	{0.258}	{0.287}
	[0.140]	[0.160]	[0.601]	[0.542]	[0.931]	[0.839]	[0.709]	[0.912]	[0.748]	[0.391]	[0.906]	[0.980]	[0.597]	[0.697]	[0.181]	[0.219]
Observations	163	160	171	168	163	160	163	160	163	160	163	160	163	160	172	169
First-wave FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓		✓		✓		✓		✓		✓

Notes: This table reports robustness results for the narrative-engagement outcomes. The six column groups correspond to: (i) *Transportation*; (ii) *Realism*; (iii) *Cognitive access*; (iv) *Loss of time*; (v) *Loss of self*; and (vi) *Narrative distraction.* All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the narrative-engagement family, and the randomization-inference p-value based on 999 permutations.

Table A.26: RI and MHT on Narrative scales

	Transportation		Realism		Cognitive access		Loss of time		Loss of self		Narrative distraction	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	(0.064)	(0.045)	(0.578)	(0.295)	(0.130)	(0.014)	(0.149)	(0.062)	(0.001)	(0.002)	(0.352)	(0.182)
	{0.193}	{0.089}	{0.578}	{0.295}	{0.223}	{0.041}	{0.223}	{0.093}	{0.008}	{0.012}	{0.423}	{0.218}
	[0.157]	[0.177]	[0.665]	[0.518]	[0.225]	[0.138]	[0.256]	[0.212]	[0.036]	[0.059]	[0.469]	[0.333]
Observations	163	160	163	160	163	160	163	160	163	160	163	160
First-wave FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓		✓		✓		✓

Notes: This table reports robustness results for the other psychological and rival channels family. The eight column groups correspond to: (i) *Transportation*; (ii) *RMET*; (iii) *Affective empathy*; (iv) *Cognitive empathy*; (v) *Moral symbolization*; (vi) *Moral internalization*; (vii) *UCLA loneliness*; and (viii) *IOS summary*. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the other psychological and rival channels family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.27: RI and MHT on Empathy scales

	Affective empathy		Cognitive empathy	
	(1)	(2)	(3)	(4)
Treatment	(0.893)	(0.764)	(0.663)	(0.890)
	{0.893}	{0.890}	{0.893}	{0.890}
	[0.939]	[0.827]	[0.718]	[0.917]
Observations	163	160	163	160
First-wave FE	✓	✓	✓	✓
Controls		✓		✓

Notes: This table reports robustness results for the empathy outcomes. The two column groups correspond to: (i) *Affective empathy*; and (ii) *Cognitive empathy*. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the RMET family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.28: RI and MHT on Uniqueness scales

	Uniqueness Orientation		Attention to individual differences	
	(1)	(2)	(3)	(4)
Treatment	(0.616)	(0.547)	(0.048)	(0.031)
	{0.616}	{0.547}	{0.095}	{0.063}
	[0.675]	[0.661]	[0.149]	[0.172]
Observations	163	160	163	160
First-wave FE	✓	✓	✓	✓
Controls		✓		✓

Notes: This table reports robustness results for the uniqueness outcomes. The two column groups correspond to: (i) *Uniqueness*; and (ii) *Uniqueness (attentional)*. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the RMET family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.29: RI and MHT on Atributtion scales

	Attribution index		Complexity		Situational		Motivation		Metacognitive		Ambiguity		Cognitive effort		Behavior	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Treatment	(0.127)	(0.096)	(0.143)	(0.130)	(0.412)	(0.428)	(0.770)	(0.734)	(0.101)	(0.033)	(0.089)	(0.036)	(0.132)	(0.015)	(0.176)	(0.237)
	{0.228}	{0.193}	{0.228}	{0.209}	{0.470}	{0.490}	{0.770}	{0.734}	{0.228}	{0.097}	{0.228}	{0.097}	{0.228}	{0.097}	{0.235}	{0.315}
	[0.159]	[0.211]	[0.213]	[0.264]	[0.551]	[0.635]	[0.856]	[0.839]	[0.158]	[0.138]	[0.175]	[0.181]	[0.165]	[0.085]	[0.223]	[0.291]
Observations	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169
First-wave FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓		✓		✓		✓		✓		✓

Notes: This table reports robustness results for the attribution outcomes. The eight column groups correspond to: (i) *Attribution index*; (ii) *Complexity*; (iii) *Situational*; (iv) *Motivation*; (v) *Metacognitive*; (vi) *Ambiguity*; (vii) *Cognitive effort*; and (viii) *Behavior*. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the RMET family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.30: RI and MHT on Morality scales

	Symbolization		Internalization	
	(1)	(2)	(3)	(4)
Treatment	(0.707)	(0.248)	(0.861)	(0.978)
	{0.861}	{0.497}	{0.861}	{0.978}
	[0.763]	[0.399]	[0.906]	[0.988]
Observations	163	160	163	160
First-wave FE	✓	✓	✓	✓
Controls		✓		✓

Notes: This table reports robustness results for the moral-identity outcomes. The two column groups correspond to: (i) *Symbolization*; and (ii) *Internalization*. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In each outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the moral-identity family, and the randomization-inference p-value based on 999 permutations.

Table A.31: RI and MHT on IOS scales

	IOS index		Family		Close friends		Studying partners		Neighbors		Higher strata		Lower strata		Different political ideas		Other nationality		Unknown person	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Treatment	(0.065)	(0.072)	(0.741)	(0.746)	(0.479)	(0.341)	(0.012)	(0.011)	(0.009)	(0.009)	(0.181)	(0.318)	(0.027)	(0.030)	(0.093)	(0.112)	(0.368)	(0.252)	(0.758)	(0.487)
	{0.161}	{0.180}	{0.758}	{0.746}	{0.599}	{0.426}	{0.058}	{0.054}	{0.058}	{0.054}	{0.301}	{0.426}	{0.091}	{0.099}	{0.186}	{0.223}	{0.526}	{0.420}	{0.758}	{0.542}
	[0.181]	[0.219]	[0.854]	[0.802]	[0.550]	[0.519]	[0.032]	[0.076]	[0.061]	[0.091]	[0.295]	[0.483]	[0.132]	[0.131]	[0.178]	[0.251]	[0.439]	[0.405]	[0.803]	[0.635]
Observations	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169	172	169
First-wave FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓

Notes: This table reports robustness results for the IOS outcomes. The ten column groups correspond to: (i) *IOS index*; (ii) *Family*; (iii) *Close friends*; (iv) *Studying partners*; (v) *Neighbors*; (vi) *Higher strata*; (vii) *Lower strata*; (viii) *Different political ideas*; (ix) *Other nationality*; and (x) *Unknown person*. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In this outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the RMET family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.32: RI and MHT on Loneliness scale

	UCLA loneliness	
	(1)	(2)
Treatment	(0.513)	(0.615)
	{0.513}	{0.615}
	[0.593]	[0.691]
Observations	163	160
First-wave FE	✓	✓
Controls		✓

Notes: This table reports robustness results for the UCLA loneliness outcome. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In this outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the RMET family in braces, and the randomization-inference p-value based on 999 permutations in brackets.

Table A.33: RI and MHT on RMET scale

	RMET	
	(1)	(2)
Treatment	(0.442)	(0.357)
	{0.442}	{0.357}
	[0.601]	[0.542]
Observations	171	168
First-wave FE	✓	✓
Controls		✓

Notes: This table reports robustness results for the Reading the Mind in the Eyes Test (RMET) outcome. All regressions include first-wave fixed effects. Columns marked in the row *Controls* additionally include the standard set of demographic and pre-treatment covariates. In this outcome, the three entries reported under *Treatment* correspond, from top to bottom, to the conventional p-value from the main specification in parentheses, the Benjamini–Hochberg multiple-hypothesis-adjusted p-value within the RMET family in braces, and the randomization-inference p-value based on 999 permutations in brackets. Because RMET enters the analysis as a single aggregate outcome, the multiple-hypothesis-adjusted p-value coincides mechanically with the conventional p-value.

A.3.5 Selection into the Analytic Sample

Table A.34: Effective Randomization

	(1)
Age	0.012
	(0.008)
Quizscore	-0.002
	(0.002)
Firstwave	-0.033
	(0.066)
Short term	0.015
	(0.054)
Knowother	0.061
	(0.068)
Female	-0.007
	(0.052)
High strata	-0.045
	(0.071)
Econ field	-0.030
	(0.071)
Placebo mean	0.512
Observations	385
R^2	0.016
P-value: controls joint test	0.515

Notes: This table reports a randomization balance test in the full experimental sample. The dependent variable is an indicator for assignment to the fiction treatment. Each coefficient comes from a linear regression of treatment assignment on baseline covariates: age, quiz score, first-wave participation, short-term treatment timing, whether the participant knew other subjects, female, high socioeconomic strata, and economics field. "Placebo mean" reports the mean of the dependent variable in the estimation sample. The row labeled "P-value: controls joint test" reports the p-value from a joint significance test of all baseline covariates. Robust standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.35: Effective Randomization

	(1)
Age	0.014*
	(0.008)
Quiz score	-0.004*
	(0.002)
First wave	0.006
	(0.073)
Short term	0.025
	(0.054)
Knows others	0.063
	(0.068)
Female	-0.017
	(0.054)
High strata	-0.052
	(0.073)
Economics field	-0.030
	(0.071)
Placebo mean	0.519
Observations	362
R^2	0.025
P-value: controls joint test	0.202

Notes: This table reports a randomization balance test in the restricted estimation sample. The dependent variable is an indicator for assignment to the fiction treatment. Each coefficient comes from a linear regression of treatment assignment on baseline covariates: age, quiz score, first-wave participation, short-term treatment timing, whether the participant knew other subjects, female, high socioeconomic strata, and economics field. The sample is restricted by outcome availability or scale completion; in particular, we apply an additional filter based on missingness in the game outcomes or scale measure. "Placebo mean" reports the mean of the dependent variable in the corresponding estimation sample. The row labeled "P-value: controls joint test" reports the p-value from a joint significance test of all baseline covariates. Robust standard errors are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.36: Attrition Analysis

	(1)	(2)	(3)
Assigned treatment	0.028	0.030	-0.228
	(0.199)	(0.052)	(0.352)
Age		-0.012*	-0.023
		(0.006)	(0.014)
Quiz score		-0.001	0.000
		(0.001)	(0.001)
First wave		-0.214**	-0.191*
		(0.073)	(0.091)
Short term		-0.094	-0.114
		(0.053)	(0.081)
Knows others		0.063**	0.064
		(0.027)	(0.045)
Female		0.063	0.027
		(0.039)	(0.060)
High strata		-0.013	0.029
		(0.057)	(0.086)
Economics field		-0.055	-0.088
		(0.041)	(0.054)
Assigned treatment × Age			0.013
			(0.017)
Assigned treatment × Quiz score			-0.003
			(0.002)
Assigned treatment × First wave			-0.044
			(0.137)
Assigned treatment × Short term			0.044
			(0.104)
Assigned treatment × Knows others			-0.011
			(0.055)
Assigned treatment × Female			0.079
			(0.076)
Assigned treatment × High strata			-0.098
			(0.131)
Assigned treatment × Economics field			0.081
			(0.075)
			(0.098)
Constant	0.751***	1.193***	1.411***
	(0.136)	(0.163)	(0.250)
Placebo mean	0.751	0.862	0.862
Observations	452	385	385
R^2	0.001	0.125	0.142
P-value: controls joint test		0.006	
P-value: interactions joint test			0.004
Session-clustered SE	✓	✓	✓

Notes: This table reports an attrition analysis in the unrestricted experimental sample. The dependent variable is an indicator equal to one if the participant attended at least three sessions and zero otherwise. Each column reports coefficients from a linear regression of this analytic-sample indicator on treatment assignment and, depending on the specification, baseline covariates and their interactions. Baseline covariates include age, quiz score, first-wave participation, short-term treatment timing, whether the participant knew other subjects, female, high socioeconomic strata, and economics field. The sample is not additionally restricted by outcome availability or scale completion. "Placebo mean" reports the mean of the dependent variable for the placebo group in the corresponding estimation sample. The rows labeled "P-value: controls joint test" and "P-value: interactions joint test" report the p-values from joint significance tests of the baseline covariates and the treatment-by-covariate interactions, respectively. Standard errors are clustered at the session level and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.37: Attrition Analysis

	(1)	(2)	(3)
Assigned treatment	0.014	0.020	-0.003
	(0.045)	(0.042)	(0.391)
Age		-0.011	-0.011
		(0.009)	(0.014)
Quiz score		-0.001	-0.000
		(0.001)	(0.001)
First wave		0.007	0.069
		(0.051)	(0.044)
Short term		-0.029	-0.032
		(0.038)	(0.060)
Knows others		0.044*	0.037
		(0.020)	(0.038)
Female		0.034	-0.025
		(0.035)	(0.038)
High strata		-0.039	-0.038
		(0.052)	(0.063)
Economics field		-0.050	-0.075
		(0.042)	(0.053)
Assigned treatment × Age			-0.000
			(0.019)
Assigned treatment × Quiz score			-0.001
			(0.002)
Assigned treatment × First wave			-0.121*
			(0.067)
Assigned treatment × Short term			0.010
			(0.071)
Assigned treatment × Knows others			0.005
			(0.048)
Assigned treatment × Female			0.116*
			(0.056)
Assigned treatment × High strata			-0.025
			(0.120)
Assigned treatment × Economics field			0.064
			(0.077)
Constant	0.920***	1.172***	1.188***
	(0.039)	(0.207)	(0.270)
Mean dep. var. (control assigned)	0.920	0.920	0.920
Observations	371	362	362
R^2	0.001	0.044	0.069
P-value: controls joint test		0.056	
P-value: interactions joint test			0.005
Session-clustered SE	✓	✓	✓

Standard errors clustered at the session level in parentheses. The dependent variable equals one if the individual participated in at least three sessions. Columns (2) and (3) include baseline controls; column (3) additionally includes interactions between assigned treatment and baseline covariates. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Notes: This table reports an attrition analysis in the restricted estimation sample. The dependent variable is an indicator equal to one if the participant attended at least three sessions and zero otherwise. Each column reports coefficients from a linear regression of this analytic-sample indicator on treatment assignment and, depending on the specification, baseline covariates and their interactions. Baseline covariates include age, quiz score, first-wave participation, short-term treatment timing, whether the participant knew other subjects, female, high socioeconomic strata, and economics field. The sample is restricted by outcome availability or scale completion; in particular, we apply an additional filter based on missingness in the game outcomes or scale measure. "Placebo mean" reports the mean of the dependent variable for the placebo group in the corresponding estimation sample. The rows labeled "P-value: controls joint test" and "P-value: interactions joint test" report the p-values from joint significance tests of the baseline covariates and the treatment-by-covariate interactions, respectively. Standard errors are clustered at the session level and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.3.6 Placebo group behavior

Table A.38: Placebo group results on games

	Dictator	Trust	Ultimatum	Third party
	(1)	(2)	(3)	(4)
Completed at least 3 sessions	-0.345	1.531*	-0.187	-0.401
	(0.712)	(0.465)	(0.093)	(0.150)
Mean (< 3 sessions)	3.875	4.250	5.000	5.000
Observations	90	90	90	90
R^2	0.101	0.203	0.104	0.095
Controls	✓	✓	✓	✓
Session-clustered SE	✓	✓	✓	✓

Notes: This table reports within-placebo correlations between course completion and observed behavior in the experimental games. The sample is restricted to participants assigned to the placebo arm, defined as those with *assigned_treatment = 0*. The main explanatory variable is an indicator equal to one if the participant attended at least three sessions of the placebo course and zero otherwise. Each column reports coefficients from a separate linear regression of game behavior on this completion indicator, controlling for baseline covariates: age, quiz score, first-wave participation, short-term treatment timing, whether the participant knew other subjects, female, high socioeconomic strata, and economics field. The omitted category is participants assigned to placebo who attended fewer than three sessions. "Mean (< 3 sessions)" reports the mean of the dependent variable for that omitted group in the corresponding estimation sample. Because completion of the placebo course was not randomly assigned, coefficients should be interpreted as correlational associations rather than causal effects. Standard errors are clustered at the session level and reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.3.7 RMET excercise

Table A.39: Treatment effects on RMET outcomes by timing

	Full sample								Short term								Long term							
	Share correct		Wrong modal		Wrong nonmodal | wrong		Categories used		Share correct		Wrong modal		Wrong nonmodal | wrong		Categories used		Share correct		Wrong modal		Wrong nonmodal | wrong		Categories used	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
Treatment	0.019	0.024	-0.016	-0.014	0.016	0.016	0.057	0.061	0.061	0.063	-0.020	-0.018	0.009	0.003	0.101	0.120*	-0.035***	-0.024**	-0.012*	-0.019**	0.032**	0.041**	-0.004	-0.002
	(0.019)	(0.023)	(0.010)	(0.009)	(0.031)	(0.027)	(0.042)	(0.049)	(0.033)	(0.038)	(0.021)	(0.018)	(0.053)	(0.039)	(0.046)	(0.047)	(0.008)	(0.009)	(0.005)	(0.006)	(0.011)	(0.016)	(0.053)	(0.057)
Constant	0.641***	0.579***	0.239***	0.375***	0.520***	0.259*	3.636***	3.709***	0.593***	0.513**	0.229***	0.435**	0.546***	0.149	3.690***	3.913***	0.695***	0.647***	0.252***	0.243**	0.487***	0.523**	3.590***	3.398***
	(0.017)	(0.098)	(0.009)	(0.077)	(0.019)	(0.129)	(0.023)	(0.386)	(0.033)	(0.109)	(0.018)	(0.084)	(0.030)	(0.144)	(0.007)	(0.400)	(0.004)	(0.156)	(0.004)	(0.091)	(0.009)	(0.178)	(0.038)	(0.554)
Placebo mean	0.644	0.644	0.241	0.241	0.520	0.520	3.620	3.620	0.593	0.593	0.229	0.229	0.546	0.546	3.690	3.690	0.703	0.703	0.254	0.254	0.490	0.490	3.541	3.541
Observations	171	168	171	168	171	168	171	168	90	89	90	89	90	89	90	89	80	78	80	78	80	78	80	78
First-wave FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓		✓

Notes: Entries report coefficients from regressions of RMET outcomes on an indicator for assignment to the fiction-reading treatment. The outcome in Columns (1)–(2), (9)–(10), and (17)–(18) is the share of correct responses across answered RMET items. Columns (3)–(4), (11)–(12), and (19)–(20) report the share of responses equal to the modal incorrect answer in the placebo group. Columns (5)–(6), (13)–(14), and (21)–(22) report the share of wrong nonmodal responses conditional on being incorrect. Columns (7)–(8), (15)–(16), and (23)–(24) report the number of distinct response categories used across RMET items. All specifications include first-wave fixed effects. Columns marked in the Controls row additionally include the full set of controls. In the full-sample specifications, this set includes the short-term indicator; in the short-term and long-term subsamples, regressions are estimated separately and therefore omit that indicator. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.3.8 Fiction Course

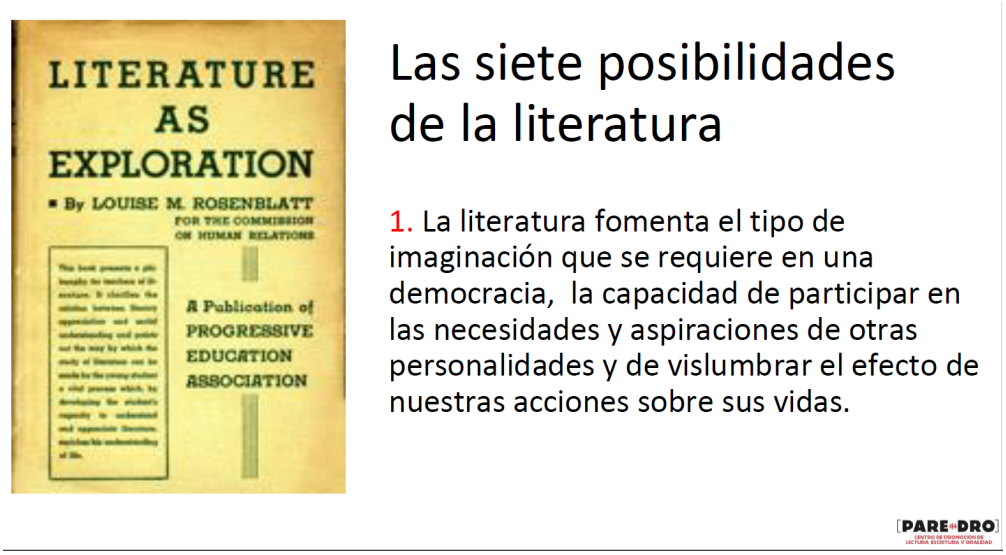


Figure A.18: Sample slide of the Fiction reading course. Session 1: The seven possibilities of literature

Figure A.19: Sample slide of the Fiction reading course. Session 3: From empathy to compassion. "Masacre en Colombia", Fernando Botero, 2000

Figure A.20: Sample slide of the Fiction reading course. Session 3: From empathy to compassion. Jesús Abad Colorado