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

Juan Ferro† José-Alberto Guerra‡

Preliminary version. Please do not circulate. For the latest version, see https://bit.ly/FG_ReadingOthers_202608

Abstract

What drives prosocial behavior, and how can it be shaped through educational interventions? We examine whether guided reading of literary fiction changes prosocial behavior. In a preregistered field experiment, university students were randomly assigned to one of two four-week extracurricular courses: guided fiction reading or behavioral science, with comparable duration and reading requirements. Following the intervention, participants completed incentivized economic games and psychological scales. We find that the fiction course increases altruism and trustworthiness. We find no robust effects on trust or costly norm-enforcement, and some prosocial norms become more permissive. While empathy does not change, we find suggestive evidence of greater narrative transportation, richer attributions about human behavior, and greater attention to individual differences, alongside lower self–other overlap. These findings are consistent with a social-cognitive mechanism centered on valuing individuation: attending to and appreciating what distinguishes people. A model of coarse reasoning illustrates how finer representations can increase average prosocial behavior without requiring stronger other-regarding preferences. Our paper contributes to the growing literature on how narrative experiences shape social behavior and suggests paths for educational interventions to increase prosociality by changing social perception, without strengthening prosocial norms.

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 and the Office of the Vice President for Research 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, Rachid Laajaj, James Robinson, and audiences at the BEEC, Universidad Javeriana, Universidad de los Andes, and TREES workshop. The filed experiment was approved by Universidad de los Andes' and Department of Economics' IRB by Acta No 88 de 2025.

†Department of Political Science, Universidad de los Andes and Secretaría de Movilidad, Alcaldía de Bogotá, jj.ferro26@uniandes.edu.co.

‡Corresponding author. Department of Economics, Universidad de los Andes, ja.guerra@uniandes.edu.co.

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 and literary studies (Brodsky, 1987; Keen, 2007; Nussbaum, 1990; Rorty, 1989; 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-analyses suggest a small positive effect of fiction reading on social cognition—an umbrella that encompasses theory of mind and empathy (Dodell-Feder & 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 randomly assign participants to a fiction-reading course (treatment) or to a parallel 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. Their parallel structure is intended to hold constant the duration and format of the educational experience, helping separate the effect of the fiction course from that of course participation itself (see Section 3). The fiction course aimed to help participants become better readers of literature. Through guided reading and discussion, participants examined narrative voice and perspective and considered how characters' actions could arise from different motives, emotions, and circumstances. The readings drew primarily on Colombian fiction and presented characters facing emotionally charged, ambiguous and morally complex situations.

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-Feder & Tamir, 2018; Green & Brock, 2000a; Kidd & Castano, 2013). We emphasize an additional channel: fiction may encourage readers to *value individuation*—to attend to and appreciate what distinguishes people—and thereby change how they represent others before making social choices. Psychological work distinguishes category-based impressions, which rely on common group representations, from individuated processing that gives greater weight to person-specific information (Fiske & Neuberg, 1990; Hilton & von Hippel, 1996). Related economic models describe coarse thinking as pooling distinct objects or situations into common categories (Fryer & Jackson, 2008; Mullainathan et al., 2008). Literary narratives are a natural setting in which such granularity may change because they repeatedly expose readers to motives, histories, ambiguities, and contradictions that make characters difficult to reduce to generic types (Castano et al., 2021, 2024). This individuation channel is conceptually distinct from self–other overlap (Aron et al., 1992): another person can become more psychologically specific without becoming more similar or closer to oneself.

Our conceptual framework formalizes this idea as a change in the *granularity* of social representation (see Section 2). A coarse social representation pools heterogeneous inferred person-specific "stories" and assigns them a common other-regarding weight, a finer representation distinguishes inferred stories that were previously treated as interchangeable. Refinement therefore need not make everyone seem better, nor raise average concern for others. Instead, it can recover heterogeneity in social valuation that categorical reasoning had pooled together. If preferences over prosocial action requires the social value of benefiting the other to be sufficiently large relative to its private costs—as in a simple threshold prosocial decision problem—this refinement can raise average prosocial behavior even though average social valuation is unchanged. The mechanism is therefore different from a uniform increase concern for others, or an updating of prevailing social norms.

Our main empirical finding, presented in Section 4, is that assignment to literary-fiction course 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 norm enforcement are salient. A second result concerns the source of these behavioral changes. Combining self-reported personal norms with incentivized elicitation of empirical and normative social expectations, we find little evidence of broad updating of perceived social norms that could account for the behavioral effects. In particular, fiction-treated students become more generous and more trustworthy even though we do not observe a general tightening of prosocial norms. At the same time, they become somewhat more permissive in some judgments of non-prosocial behavior.

Our mechanism results in Section 5 provide suggestive evidence consistent with the representation-refinement interpretation, while also helping distinguish it from several alternatives. The fiction-reading course increases narrative immersion and transportation,

showing that the intervention changed how participants engage with narrative material, but we find no robust change in affective or cognitive empathy. Treated participants also endorse richer, more multifaceted explanations of human behavior on some dimensions of the Attributional Complexity Scale and show greater attention to individual differences on the narrower component of our Perceived Uniqueness scale, although the broad Uniqueness Orientation index does not change significantly. Finally, average self–other overlap declines, with the largest estimated reductions occurring for some socially meaningful but less intimate groups. Taken together, these outcomes are consistent with an orientation toward valuing individuation without greater assimilation of others to the self.

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 incentive-compatible coordination-based method of Krupka and Weber (2013). This combination allows us to examine whether behavioral changes are accompanied by changes in personal norms, empirical expectations, or normative expectations.

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-Feder & Tamir, 2018). In our data, the behavioral effects remain significant even at the four-week follow-up. The estimated effect on attention to individual differences is also similar across horizons. This temporal pattern reinforces the value of designs that extend measurement beyond the intervention window.

Recent work first emphasizes that prosocial behavior depends on what decision-makers know and believe about others. Individuating information can weaken reliance on group-based representations and thereby alter prosocial behavior—even reducing generosity toward favored in-group members (Lee et al., 2021). Beliefs about effort, luck, and deservingness similarly shape generosity and support for redistribution (Alesina et al., 2018; Stantcheva, 2022). Consistent with this idea, Alan et al. (2023) show that strengthening meritocratic beliefs increases effort while reducing altruism toward individuals perceived as responsible for their outcomes. Information may also operate through 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 the effect operates through increased perceived similarity: individuals become more empathetic when they experience others' hardship as if they were "in their shoes." Such effects need not be uniformly prosocial, however: exposure and experience can also reinforce perceived differences and increase xenophobia (Braun, 2022). These findings relate to a broader literature on the boundaries of moral concern. Enke (2025) conceptualizes moral universalism as the degree to which

a person's altruism is invariant to the identity of the recipient. He documents systematic heterogeneity in how broadly individuals extend moral concern beyond their in-groups and relates this variation to historical factors. This perspective shows that prosociality depends not only on how much people care about others, but also on whom they care about.

A second strand examines whether educational experiences can produce lasting changes in cognitive skills and social behavior. Hoxby (2022) emphasizes the role of education in developing advanced cognitive skills during formative stages of development. Other educational interventions seek to shape social beliefs and behavior from an early age (Cappelen et al., 2020), including through curricula designed to develop perspective-taking and prosocial values. Historical and educational evidence further suggests that sustained exposure to moral narratives can produce persistent changes in values and social attitudes (Amaya et al., 2024; Figueroa & Fouka, 2023; Hartinger et al., 2026). Together, these studies demonstrate the potential of education to shape social behavior and motivate closer examination of the channels through which these effects arise.

Our contribution brings together these two strands. Like the first, we emphasize that prosocial behavior depends on how the person affected by a decision is perceived and represented. Unlike accounts centered on deservingness, perceived similarity, or the scope of moral concern, we examine whether fiction encourages readers to see others as distinct individuals rather than as interchangeable members of a category. This idea echoes Nagel (1970), who argues 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." Like the second strand, we treat the literature course as an educational intervention that may shape social behavior. By pairing the curriculum with incentivized measures of prosocial behavior and social perception, we bridge insights from the humanities and behavioral science and examine whether literary education can change how people perceive and treat others.

2 Conceptual framework

We use a minimal model to organize the interpretation of the experiment (see Appendix A.1). The model is not intended to imply that literary fiction necessarily changes social representation; rather, it characterizes the behavioral consequences if the intervention makes the representation of other people finer. The central distinction is between the underlying person-specific features that could matter for social valuation and the granularity with which those features are represented when a social choice is made.

Let Ω be a finite set of inferred *social stories*, with every story $\omega \in \Omega$ distributed according to μ. A social story is not an objectively known account of another person, nor need it correspond to that person's true characteristics. Rather, it is a possible person-specific representation held by the decision-maker: a configuration of believed or inferred motives, circumstances, histories, intentions, traits, and other features that may matter for social

valuation. Accordingly, μ may be interpreted as the decision-maker's subjective distribution over these possible representations rather than as an objective population distribution.[1] Let $v(\omega)$ denote the social valuation associated with story ω when the distinctions contained in that representation are behaviorally available. This valuation is itself subjective: it captures how much weight the decision-maker would place on the other person's welfare under that representation, not the person's objective moral worth.

A *social representation* is a partition Π of Ω. Stories in the same cell are treated as interchangeable for purposes of social valuation. When several stories are pooled, we assume that their common other-regarding weight is

$$\beta_\Pi(\omega) \equiv \mathbb{E}_\mu[v(\tilde{\omega}) \mid \tilde{\omega} \in \Pi(\omega)]\,.$$

Thus, a coarse representation compresses person-specific differences by assigning stories within the same category a common social weight.[2]

Consider a coarse representation Π_C and a finer representation Π_F, where Π_F refines Π_C. We interpret this social refinement as *individuation*: the finer representation distinguishes inferred stories that the coarse representation treated as interchangeable, without changing the underlying distribution of inferred stories or their valuations. Let $\beta_C \equiv \beta_{\Pi_C}$ and $\beta_F \equiv \beta_{\Pi_F}$. Social refinement implies (see Proposition 1)

$$\beta_C = \mathbb{E}[\beta_F \mid \Pi_C], \qquad \mathbb{E}[\beta_F] = \mathbb{E}[\beta_C], \qquad \text{and } \operatorname{Var}(\beta_F) \geq \operatorname{Var}(\beta_C) \tag{1}$$

Hence, a finer representation generates a mean-preserving spread in social weights: distinctions that were previously averaged away can raise the weight placed on some social stories and lower it on others, while leaving average social valuation unchanged.

To connect individuation to behavior, suppose the decision-maker chooses a prosocial action $a \geq 0$ that produces a units of benefit for the other person. Let utility be

$$u(a;\beta) = \beta a - \kappa a - \frac{1}{2}a^2, \qquad \kappa \geq 0.$$

The parameter κ captures the marginal private cost of prosocial action at zero. Positive action therefore requires the social value of benefiting the other, β, to be sufficiently large relative to its private cost, a reduced-form trade-off consistent with classic cost–reward accounts of helping (Dovidio, 1984).[3] The optimal prosocial action is thus

$$a^*(\beta) = \max\{\beta - \kappa, 0\}.$$

[1] The argument requires only that the same weighting be used consistently across coarse and fine representations.

[2] The conditional-mean rule need not be interpreted as literal mental averaging. When stories within a category remain unresolved and social valuation enters utility linearly, expected utility depends only on the conditional mean of that valuation.

[3] In allocation settings such as dictator-game transfers or trust-game returns, this private cost also has a literal opportunity-cost interpretation: increasing the other's payoff requires sacrificing one's own payoff.

which, combining with (1), gives (see Proposition 2)

$$\mathbb{E}[a^*(\beta_F)] \geq \mathbb{E}[a^*(\beta_C)].$$

Thus finer representation can increase expected prosocial action even though average social valuation does not change.

The model points to some empirical patterns that are informative about the proposed individuation mechanism. First, one would expect treatment effects on costly prosocial choices. Second, evidence of greater attention to person-specific differences or richer, less reductive explanations of others is consistent with finer social representation; in our data, the Perceived Uniqueness measures and Attributional Complexity Scale speak most directly to this margin. Third, refinement does not require higher average other-regarding weights or self-other overlap: representing another person more distinctly need not increase perceived similarity or closeness. Finally, the mechanism operates through individual representation and the other-regarding weights entering choice, so it need not be accompanied by changes in empirical or normative expectations. These implications motivate the mechanism analysis in Section 5, while leaving open alternative channels that the experiment can only distinguish imperfectly.

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 [4]. 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, most importantly, selection.

The experiment was conducted in two waves at Universidad de los Andes (Bogotá, Colombia), with a total of 391 students attending at least one course session. The first wave ran from late April to May 2025 and drew participants from the Uniandes Mobile Experimental Lab while the second wave ran from September to November 2025 and was open to undergraduates across the university.Both waves offered four 90-minute course sessions. Before randomization, registered students were invited to complete an entrance survey covering demographic characteristics, reading habits, and beliefs about the effects of fiction reading on attitudes. Students were randomly assigned to the literary-fiction course or the behavioral-science placebo course.

[4]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 B.1 shows the actual email sent

3.1 Course design

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[5]. The possible reading material for a 4-week course is endless. We necessarily had to chose, and we chose books that had the potential to generate a deep narrative experience. We are not arguing that any curriculum of fiction readings can increase 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[6] is used as a canonical conception of what literature *can* do. The first of the seven possibilities discussed is the creation of the necessary moral imagination that democracy need in its citizens 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. 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

[5]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. This connects with the idea expressed by Rorty (1989), that solidarity is "created by increasing our sensitivity to the particular details of the pain and humiliation of other, unfamiliar sorts o people. Such increased sensitivity makes it more difficult to marginalize people different from ourselves(...)"

[6]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)

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 Economics Ph.D.'s. 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 Bursztyn et al. (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 Online 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 in both the treatment and placebo course completed an incentivized quiz on the course reading material (to encourage thorough engagement) with equal monetary incentives. 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.

3.2 Incentivized experimental games and norm elicitation

Participants completed a set of incentivized economic games capturing 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 where 1 token was equivalent to roughly USD 0.5. Participants in both treatment conditions were compensated if they participated in at least three course sessions (USD 22 approx.), and they could earn additionally a maximum of USD 16.5 for their decisions in each of the experimental sessions.

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.[7]

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 back, and Player 2 beliefs about the most common amount sent by trustors.

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.

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.

Finally, 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 2 tokens if their response matched what other random chosen participant would answer for the same question.

[7] All empirical expectations in the experimental games were incentivized with a payment of 2 tokens for an exactly correct response.

3.3 Psychological scales

To shed light on the mechanisms underlying the behavioral effects, participants also completed a set of psychological measures capturing narrative engagement and several dimensions of social cognition. We measured narrative immersion using items based on the Transportation framework (Green & Brock, 2000b) and the Narrative Engagement Scale (Busselle & Bilandzic, 2009), perceived self–other overlap using the Inclusion of Other in the Self (IOS) scale (Aron et al., 1992), the tendency to construct multifactorial explanations of others' behavior using the Attributional Complexity Scale (Fletcher et al., 1986), moral self-concept using the Moral Identity Scale (Aquino & Reed, 2002), affective theory of mind using the revised Reading the Mind in the Eyes Test (RMET) (Baron-Cohen et al., 2001), and construct narrative-based measures of affective and cognitive empathy from the post-reading battery.[8]

Because existing instruments do not directly capture the individuation mechanism from our conceptual framework, we additionally developed a Perceived Uniqueness scale, informed by the distinction between categorical and individuated processing (Fiske & Neuberg, 1990). The scale measures the extent to which participants attend to person-specific differences and resist treating individuals as interchangeable members of social categories. We describe the construction and scoring of all measures in Appendix B.4 and discuss their treatment effects in Section 5.

The administration of these games and psychological scales were organized into two sets. Set A comprised the dictator and trust games, together with measures of perceived uniqueness, perceived social isolation, narrative engagement, and moral identity. The narrative-based empathy measures were constructed from the post-reading responses collected in this set. Set B comprised the ultimatum and third-party punishment games, together with measures of attributional complexity, self–other overlap, and theory of mind.

Participants were randomly assigned to the order A→B or B→A, with one set of measures scheduled for the short-term experimental session and the other for the longer-term session. Short-term sessions were held immediately after the final course session or one week later the longer-term session was held approximately four weeks after the course ended.[9]

4 Main results

In this section we present the main empirical results. Experimental outcomes are available for 248 students. Table 1 describes 247 of these students, excluding one whose Author Recognition Test score is missing. Sample sizes vary across outcome regressions according to

[8] We also measured perceived social isolation using the UCLA Loneliness Scale (Russell, 1996).

[9] The first wave followed a different measurement schedule: all economic games were administered immediately after the final course session, while the psychological measures were collected approximately four months later.

the availability of each measure. Course participation and experimental outcome availability by treatment assignment are examined in Online Appendix B.5.1.

Table 1: Sample balance. Demographics.

	(1) Mean	(2) Sd	(3) Treated	(4) Placebo	(5) pval
Age	20.11	2.72	20.32	19.88	0.196
Quiz score	19.58	13.06	17.87	21.59	0.030
First wave	0.15	0.36	0.16	0.17	0.831
Only short term	0.45	0.50	0.46	0.42	0.555
Know other	0.17	0.37	0.19	0.17	0.699
Female	0.60	0.49	0.58	0.58	0.913
High strata	0.16	0.37	0.17	0.18	0.835
Econ field	0.17	0.38	0.18	0.18	0.919
Fiction read	3.16	4.05	3.29	2.89	0.450
Observations			133	114	

Notes: This table compares pretreatment characteristics and follow-up participation by treatment assignment among students with at least one observed experimental game outcome or psychological measure. Of the 248 students with experimental data, one is excluded because their Author Recognition Test score is missing, leaving 247 students in the table. Columns (1)–(2) show the overall mean and standard deviation in the full sample. The sample is restricted to individuals with non-missing information for at least one experimental game outcome or one scale measure. Columns (3)–(4) report group means by assignment status. Column (5) reports two-sided p-values from robust regressions testing 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 only short-term sessions, whether the participant was acquainted with someone from the opposite treatment condition, 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, as the total of writers (from a maxmimum of 45) that where correctly guessed. Fiction read denotes the number of fiction books read.

The observed sample is relatively young, consistent with the university recruitment setting. The baseline comparisons show broadly similar observed characteristics across treatment arms. The treatment group reports slightly higher average reading hours, but not a higher number of books read or purchased during the previous year.

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.[10] As the table shows, the placebo group scores higher on this measure, 21.59 compared with 17.87 in the treatment group ($p = 0.030$). This difference does not support the possibility that treated participants 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 B.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 slightly more inclined to associate fiction with greater psychological complexity than with empathy per se, but in our sample fiction is not seeing as a way of increasing pos-social behavior.

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

$$y_{i,t} = \alpha + \beta \, \text{Treatment}_i + \pi_{w(i)} + \mathbf{x}_i' \boldsymbol{\gamma} + \varepsilon_{i,t}, \tag{2}$$

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 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 transferred in the third-party punishment game. α is a constant term. Treatment_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 estimation sample includes participants with least one experimental session.

Course assignment was concealed until participants arrived, and both arms received equivalent invitations, schedules, and reminders. Initial attendance therefore occurred before participants learned which course they would receive. The coefficient of interest, β, therefore captures the average treatment effect of fiction exposure on the corresponding behavioral outcome.[11] $\pi_{w(i)}$ 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 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

[10] As shown in Figure B.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.

[11] Its interpretation as an effect of assignment depends on whether selection into the observed-outcome sample introduces bias, we examine participation patterns in Online Appendix B.5.1.

participant across the relevant within-game actions of Player 1 and estimate:

$$y_{i,s,t} = \alpha + \beta\,\text{Treatment}_i + \sum_{k \neq s_0} \delta_k \mathbf{1}\{s = k\} + \sum_{k \neq s_0} \theta_k\,(\text{Treatment}_i \times \mathbf{1}\{s = k\}) + \pi_{w(i)} + \mathbf{x}_i'\boldsymbol{\gamma} + \varepsilon_{i,s,t}. \tag{3}$$

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

The index s refers to the strategy-method potential behavior of Player 1, while k indexes the set of possible actions included in the summation. Thus, $\mathbf{1}\{s = k\}$ is an indicator equal to one when the observed decision corresponds to Player 1 sent amount s. The omitted contingency s_0 is an omitted reference dependent on the relevant game. In the trust game, the outcome is the share of the received amount returned by the trustee for each positive transfer by Player 1. In the ultimatum game, it is the acceptance decision for each possible offer. In the third-party punishment game, it is the punishment chosen for each allocation.

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

The results reveal that exposure to a guided-literary fiction course 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.

4.1 Altruism

The dictator game provides the cleanest behavioral test on prosocial action as giving is costly to the decision-maker, directly benefits another participant, and does not depend on the recipient's subsequent action. Tables 2 and 3 jointly examine whether the treatment changed this choice and whether any change was accompanied by corresponding changes in beliefs or social norms.

Table 2 reports Player 1 transfers (columns 1 and 2) alongside incentivized empirical expectations about the typical transfer (columns 3 and 4). Participants in the placebo group give, on average, 3.64 of the 10 available tokens. Assignment to the fiction-reading course increases giving by 0.72 tokens in the specification with wave fixed effects and by 0.67 tokens after adding the baseline controls. The latter corresponds to an increase of 6.71 percentage points of the initial endowment, or approximately 18.5% relative to the placebo mean. The similarity of the estimates across both specifications is also reassuring: the result is not sensitive to the inclusion of baseline covariates.

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

	Amount sent		Expectations	
	(1)	(2)	(3)	(4)
Treatment	0.716**	0.671**	0.361	0.387**
	(0.265)	(0.232)	(0.277)	(0.166)
Constant	3.637***	3.518***	3.517***	2.921***
	(0.144)	(0.783)	(0.246)	(0.882)
Placebo mean	3.637	3.637	3.516	3.516
Observations	190	190	190	190
R^2	0.037	0.118	0.024	0.128
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

The same table shows some changes in empirical expectations, although the evidence is

weaker. Relative to a placebo mean of 3.52 tokens, the estimated treatment effect on the expected modal transfer is 0.36 tokens without controls and 0.39 tokens with controls. The controlled estimate corresponds to an increase of approximately 11% relative to the placebo mean and is significant at the 5% level. Thus, increased giving is accompanied by some evidence of higher expectations about what others typically give.

Table 3 turns to norms. Participants report both their own personal norm of the appropriateness of giving less than half of the endowment and an incentivized normative expectation about how others evaluate the same action, on a 1–4 scale in which higher values indicate greater appropriateness. In the placebo group, the mean personal-norm rating is 2.65, slightly above the scale midpoint, whereas the mean normative expectation is 2.17, indicating that participants expect others to evaluate giving less than half somewhat less socially favorably than they do themselves. The treatment coefficients are small and imprecisely estimated: -0.04 for personal norms and -0.06 for normative expectations in the specifications with controls. Thus, the increase in dictator giving is not accompanied by a comparably clear shift in either participants' own personal norm or their normative social expectations, and can be attributted to a shift in their preferences.

Table 3: Player 2 Personal and social norms regarding selfishness

	Personal norm		Normative expectations	
	(1)	(2)	(3)	(4)
Treatment	-0.033	-0.038	-0.034	-0.055
	(0.094)	(0.095)	(0.110)	(0.114)
Constant	2.649***	2.538***	2.165***	2.135***
	(0.077)	(0.280)	(0.088)	(0.286)
Placebo mean	2.648	2.648	2.165	2.165
Observations	190	190	190	190
R^2	0.019	0.045	0.007	0.046
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Taken together, the dictator game results suggest that fiction increases costly transfers to

another person, while changes in empirical expectations is weaker accompanied with no change in normative expectations.

4.2 Trust and trustworthiness

The trust game separates two distinct behavioral margins. Player 1's transfer captures trust, which depends partly on beliefs about how Player 2 will respond. Player 2's return decision captures trustworthiness and allows us to distinguish a general willingness to return money from conditional reciprocity, that is, whether returns increase with the amount initially sent by Player 1.

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

	Amount sent		Expectations	
	(1)	(2)	(3)	(4)
Treatment	-0.196	-0.243	0.054	0.065
	(0.381)	(0.348)	(0.284)	(0.632)
Constant	6.276***	9.411***	5.440***	5.061
	(0.275)	(1.559)	(0.224)	(6.397)
Placebo mean	6.275	6.275	5.440	5.440
Observations	190	190	190	190
R^2	0.014	0.056	0.004	0.060
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4 first reports Player 1 behavior and empirical expectations. Participants in the placebo group send on average 6.28 of the 10 available tokens. Assignment to the fiction-reading treatment changes this amount by -0.24 tokens in the specification with controls, an estimate that is small relative to the placebo mean and non-significant. Empirical expectations show a similar pattern: the placebo mean is 5.44 tokens and the controlled treatment estimate is

0.065. Thus, unlike in the dictator game, we find little evidence that the treatment increases either trusting behavior or expectations about how much others will trust.

Table 5 reports participants' own assessment of sending fewer than five tokens and their incentivized normative expectations about the same action. Both are measured on a 1–4 scale, with higher values indicating greater acceptance of low-trust behavior. In the placebo group, the corresponding means are 2.12 and 1.90. The treatment increases these measures by 0.22 and 0.12 points, respectively, in the specifications with controls. These results do not suggest that the treatment strengthened a norm in favor of trust. If anything, treated participants become somewhat more permissive toward low-trust behavior.

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

	Personal norm		Normative expectations	
	(1)	(2)	(3)	(4)
Treatment	0.233**	0.217***	0.159**	0.122*
	(0.075)	(0.058)	(0.071)	(0.056)
Constant	2.121***	2.601***	1.901***	1.701***
	(0.033)	(0.429)	(0.013)	(0.233)
Placebo mean	2.121	2.121	1.901	1.901
Observations	190	190	190	190
R^2	0.019	0.055	0.016	0.060
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

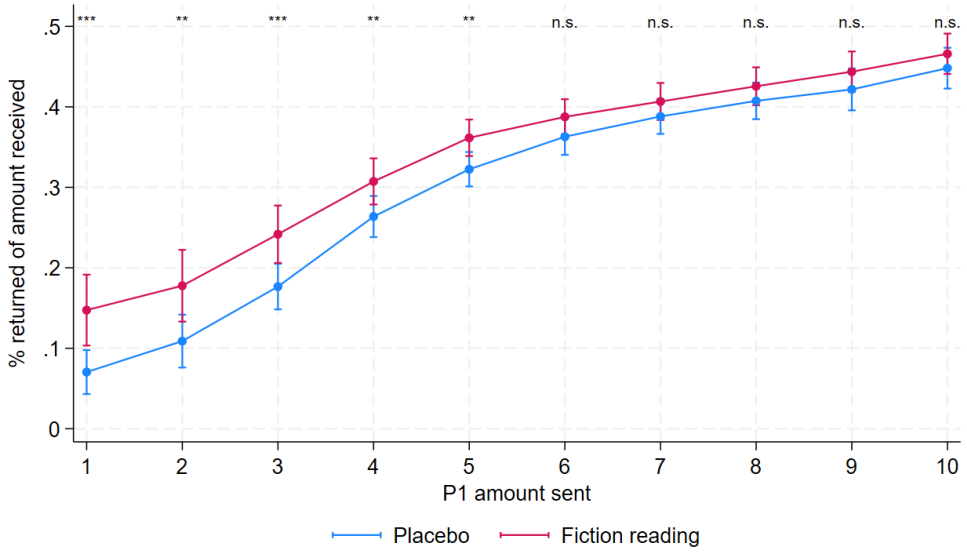
Notes: Entries report coefficients from linear regressions of trust-game norm-related outcomes on an indicator for assignment to the fiction-reading treatment. The dependent variables are (i) participants' personal norm regarding sending fewer than five tokens in the trust game, and (ii) perceived social 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Player 2 behavior provides the main trust-game result. Figure 1 plots the share of the amount received that trustees return for each possible transfer by Player 1. Fiction-treated participants return a significantly larger share when Player 1 sends half or less of the endowment. The estimated differences remain positive but are smaller and statistically insignificant for transfers of six tokens or more. The corresponding linear-interaction

specification also yields a negative interaction between treatment assignment and the amount sent, suggesting that the treatment–placebo difference narrows as transfers increase. This interaction is significant at the 10% level.[12]

This pattern is more consistent with greater trustworthiness than with stronger conditional reciprocity. Importantly, we do not observe a comparable shift in participants' personal norms or normative expectations regarding low trustee returns.[13] The trust game sharpens the pattern found in the dictator game: the treatment does not increase the willingness to place money at risk on another person's response, but it does increase prosocial behavior when the participant decides how much to return to another person. As in the dictator game, this behavioral change is not accompanied by a general strengthening of prosocial norms.

Figure 1: Share of the amount received returned by trustees, 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.

[12]The corresponding regression estimates are reported in Online Appendix Table B.12.

[13]See Online Appendix Table B.11.

4.3 Strategic fairness and norm enforcement

We find little evidence of systematic treatment effects in the ultimatum game.[14] Neither proposer offers nor responder acceptance decisions differ clearly across treatment conditions. Unlike the dictator game, ultimatum choices combine distributive preferences with strategic considerations about the other player's response. The absence of clear treatment effects is therefore consistent with the broader pattern of stronger effects on altruism and trustworthiness than on strategically contingent behavior.

The third-party punishment game yields a similarly limited behavioral response.[15] We find no robust treatment effect on Player 1 allocations or on punishment by Player 3 in response to unfair allocations, and little evidence of systematic changes in empirical or normative expectations.

Taken together with the results from the dictator and trust games, these findings provide little evidence that the fiction course changed strategic behavior or strengthened norm enforcement. Instead, the treatment increased altruism and trustworthiness while leaving behavior in games involving strategic responses or punishment largely unchanged. The positive effects on giving and trustworthiness coexist with greater permissiveness toward low trust. One possible interpretation, consistent with richer representations of others, is that treated participants may be more willing to allow for different motives or circumstances when evaluating departures from a prosocial norm. This interpretation does not require a change in expectations about others' strategic behavior, but rather a change in how others' behavior is interpreted once observed.

5 Mechanisms and robustness

We next examine how the fiction course changed participants' engagement with literary narratives and their perception of other people. The psychological measures play different roles. Narrative engagement allows us to assess whether the course changed reading in the direction intended by the curriculum. Perceived Uniqueness, Attributional Complexity, and self–other overlap speak to how participants perceive and differentiate others. We then consider empathy and theory of mind, two prominent alternative channels in the fiction-reading literature.

For each scale-based outcome, we estimate the analogue of Equation (2), including wave fixed effects and the same baseline controls used in the behavioral specifications.

[14] See Online Appendix B.5.3.
[15] See Online Appendix B.5.4.

5.1 Narrative engagement

A first question is whether the course changed how participants read. The fiction curriculum was designed not simply to expose participants to literary texts, but to make them more engaged readers: to attend closely to narrative detail, enter the narrated world, and sustain attention to characters, motives, and circumstances. To measure this outcome, participants completed a 47-item narrative-experience battery after reading Horacio Quiroga's *"El hijo"* (see Table B.5 for the wording of the items). The instrument draws on the Transportation framework of Green and Brock (2000b) and the Narrative Engagement Scale of Busselle and Bilandzic (2009). In the main text, we focus on Transportation, Cognitive access, Presence, and Involvement, four dimensions that capture absorption in the narrative and sustained engagement with the narrated world. Their internal consistency in our sample ranges from Cronbach's $\alpha = 0.65$ to $\alpha = 0.86$. The remaining subscales are reported in Online Appendix B.5.9.

The results indicate that the course changed reading in the intended direction. As shown in Table 6, treated participants score between 0.23 and 0.40 points higher on these dimensions in the specifications with controls, on a 1–7 response scale. They report greater transportation, greater cognitive access to the story, a stronger sense of narrative presence, and greater involvement. The broader battery also shows increases in loss of time and loss of self-awareness while reading in the controlled specifications.

These results suggest that the intervention changed how participants engaged with literary material. Sustained attention to characters' motives, circumstances, and perspectives was the reading practice through which the course sought to develop moral imagination. The findings are consistent with stronger engagement in that practice, although they are not a direct measure of moral imagination itself.

5.2 Individuation and the perception of others

Section 2 formalizes individuation as a refinement of social representation, distinctions between inferred person-specific social stories that are pooled under a coarse representation become separately represented. The psychological measures in this subsection provide reduced-form evidence on different aspects of that representational assumption. They are not structural measures of the partition Π or of the resulting social weights β. Rather, they allow us to ask whether the fiction course made participants more attentive to individual differences, more willing to explain behavior through multiple motives and circumstances, and less likely to perceive others as overlapping with the self.

Table 6: Narrative engagement: selected dimensions

	Transportation		Cognitive access		Presence		Involvement	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.307**	0.377**	0.234*	0.313***	0.341***	0.397***	0.164*	0.228**
	(0.132)	(0.139)	(0.115)	(0.090)	(0.089)	(0.103)	(0.083)	(0.084)
Constant	4.060***	3.456***	4.943***	4.437***	4.830***	3.814***	5.198***	3.456***
	(0.050)	(0.477)	(0.060)	(0.682)	(0.063)	(0.415)	(0.068)	(0.489)
Placebo mean	4.055	4.055	4.945	4.945	4.824	4.824	5.188	5.188
Observations	175	175	175	175	175	175	175	175
R^2	0.063	0.123	0.016	0.052	0.078	0.132	0.117	0.178
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 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, *Cognitive access* = mean of P34–P36, *Presence* = mean of P21–P24, and *Involvement* = mean of P25–P30. See Appendix Table B.5 for the full list of items and coding details. The remaining narrative subscales are reported in Online Appendix Table B.38. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.2.1 Perceived uniqueness

To capture *individuation*—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 and 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. Appendix Table B.3 reports the full item wording and coding.

We pre-specified two measures based on a principal components analysis. The broad *Uniqueness Orientation* index is the respondent-level mean of all ten items after reverse-coding (Cronbach's $\alpha = 0.77$). Higher values indicate a stronger tendency to perceive others as psychologically distinct individuals rather than as interchangeable members of categories. The narrower *Attention to Individual Differences* index is a two-item subscale (Cronbach's $\alpha = 0.72$) that isolates explicit attention and valuation to the features that distinguish one person from another.

In the pooled sample (Table 7), we do not find a statistically significant treatment effect on the broader Uniqueness Orientation index. The narrower Attention to Individual

Differences measure, however, increases by 0.25 points and remains statistically significant at conventional levels in the specification with controls. This measure maps directly into the representational distinction in Section 2: greater attention to inferred person-specific features makes distinctions that might otherwise be pooled within a common social category more salient when forming impressions of others.

Table 7: Perceived Uniqueness Indices

	Uniqueness Orientation		Attention to Individual Differences	
	(1)	(2)	(3)	(4)
Treatment	0.042	0.057	0.225*	0.246**
	(0.066)	(0.058)	(0.109)	(0.099)
Constant	3.883***	3.554***	3.547***	3.814***
	(0.047)	(0.364)	(0.064)	(0.366)
Placebo mean	3.882	3.882	3.547	3.547
Observations	175	175	175	175
R^2	0.002	0.023	0.015	0.049
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 B.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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5.2.2 Attributional complexity

The conceptual framework is related not only to whether individuals are distinguished from one another, but also to the richness of the inferred person-specific motives, circumstances, and characteristics through which others are represented. We therefore examine whether the course changed how participants explain human behavior using the 28-item Attributional Complexity Scale of Fletcher et al. (1986) (see appendix Table B.7 for the full list of

items and coding details.). The scale measures the tendency to seek multiple, context-dependent explanations of behavior rather than rely on simple attributions. After reverse-coding negatively keyed items, we construct an overall index as the respondent-level mean across the 28 items. The overall scale has high internal consistency in our sample (Cronbach's $\alpha = 0.89$).

The original instrument also distinguishes seven subscales. In the main text, we report the overall index together with three dimensions that speak most directly to the richness of inferred explanations of others' behavior: *Complex*, *Ambiguity*, and *Cognitive effort*. The complete seven-subscale results are reported in Online Appendix Table B.39.

As shown in Table 8, the treatment increases the overall Attributional Complexity index by 0.17 points in the specification with controls, although this estimate is not statistically significant. The effects are clearer for the selected dimensions: treated participants are more likely to endorse complex explanations of behavior and to recognize ambiguity in human motivations. In the language of Section 2, these results are consistent with less compressed representations of others. Rather than reducing another person's behavior to a single trait or cause, participants appear more willing to represent it through multiple motives, circumstances, and possible explanations.

Table 8: Attributional complexity: selected dimensions

	Attribution (mean)		Complex		Ambiguity		Cognitive effort	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.153	0.155	0.134	0.177**	0.179*	0.188*	0.167	0.165
	(0.087)	(0.095)	(0.076)	(0.059)	(0.084)	(0.089)	(0.112)	(0.128)
Constant	5.018***	4.906***	5.208***	5.662***	4.884***	5.189***	4.804***	4.816***
	(0.035)	(0.504)	(0.055)	(0.384)	(0.044)	(0.614)	(0.048)	(0.592)
Placebo mean	5.022	5.022	5.212	5.212	4.885	4.885	4.806	4.806
Observations	184	184	184	184	184	184	184	184
R^2	0.029	0.098	0.015	0.093	0.013	0.054	0.010	0.096
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 the 28 attribution items. Negatively keyed items are reverse-coded so that higher values consistently indicate greater attributional complexity. Columns (3)–(8) report three attributional complexity subscales: *Complex* (items 2, 9, 16, and 23), *Ambiguity* (items 5, 12, 19, and 26), and *Cognitive effort* (items 7, 14, 21, and 28). Each index is computed as a respondent-level mean over non-missing items. 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 B.7 for the full list of items and coding details. Results for all seven subscales are reported in Online Appendix Table B.39.

5.2.3 Self-other overlap

A finer representation of another person need not make that person feel more similar or closer to the self. This distinction is important for the mechanism in Section 2: individuation is related to how finely another person is represented, not to whether that person is assimilated to the decision-maker. We examine this distinction using the Inclusion of Other in the Self (IOS) scale (Aron et al., 1992). Participants reported their perceived closeness to nine groups of people, ranging from family and close friends to study partners, neighbors, people from different socioeconomic or political backgrounds, people of another nationality, and an unknown person. For each group, they selected one of seven pairs of increasingly overlapping circles, with higher values indicating greater perceived overlap between self and other.

The treatment lowers the overall IOS index by 0.36 points in the specification with controls. The group-specific estimates suggest that this decline is not uniform across relationships. Table 9 reports all nine groups, ordered by their average IOS score in the placebo group. The largest estimated reductions occur for people from lower socioeconomic strata, neighbors, and study partners. There is weaker evidence of reduced overlap with close friends and people holding different political views, with both estimates significant at the 10% level. The estimates for family and an unknown person are not statistically significant. Descriptively, the largest reductions therefore occur for some socially meaningful but less intimate groups

One interpretation is that the scope for representational refinement differs across these relationships. Close others may already be represented using rich, person-specific information, leaving relatively little room for the course to change how they are perceived. An unknown person, by contrast, provides too little information for much person-specific inference. For socially meaningful but less intimate others—including study partners, which are particularly relevant in our student population—there may be greater scope for the course to make inferred motives, circumstances, and individual differences behaviorally salient. In this sense, individuation need not bring others closer to the self. A finer representation may instead reveal ways in which another person differs from oneself, reducing the tendency to represent that person as simply "like me." The IOS results are consistent with this form of differentiation: for some groups, treated participants appear to perceive others as more distinct from themselves rather than more assimilated to the self.

5.3 Alternative social-cognitive channels

Empathy is one of the most prominent mechanisms proposed in the literature on fiction reading. From the post-story battery, we construct separate affective and cognitive empathy indices. The affective index captures emotional resonance with the story characters (Cronbach's $\alpha = 0.83$), while the cognitive index captures perspective-taking and understanding of characters' mental states (Cronbach's $\alpha = 0.81$). Higher values indicate stronger empathic responses within the narrative context.

Table 9: Self–other overlap across groups

	Placebo mean	Baseline	With controls
IOS Index (mean)	3.616	-0.379**	-0.357**
		(0.127)	(0.133)
Family	5.341	-0.176	-0.154
		(0.189)	(0.192)
Close friends	5.165	-0.262*	-0.284*
		(0.141)	(0.144)
Study partners	4.141	-0.565***	-0.531***
		(0.142)	(0.150)
Lower strata	3.741	-0.658***	-0.663***
		(0.180)	(0.183)
Higher strata	3.482	-0.421*	-0.315
		(0.222)	(0.232)
Other nationality	3.282	-0.205	-0.166
		(0.139)	(0.124)
Neighbors	2.988	-0.675***	-0.650***
		(0.184)	(0.177)
Different political ideas	2.588	-0.366	-0.389*
		(0.205)	(0.212)
Unknown person	1.812	-0.082	-0.066
		(0.155)	(0.156)
Observations		184	184
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 report their perceived closeness to nine groups of people by choosing, for each group, one of seven pairs of increasingly overlapping circles, with higher values indicating greater perceived overlap between self and other. The IOS index is constructed as the respondent-level mean across the nine measures. Groups are ordered from highest to lowest according to their average IOS score in the placebo group. Baseline specifications include fixed effects for the first experimental wave. Specifications 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. Standard errors clustered at the treatment session term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The original column-by-column specification is reported in Online Appendix Table B.40.

As shown in Table 10, we find little evidence that the fiction course changed either affective or cognitive empathy. The controlled treatment estimates are close to zero for both outcomes. Thus, although treated participants engage more deeply with literary narratives, this greater engagement is not accompanied by a detectable increase in emotional sharing or narrative-based perspective-taking as measured by these indices.[16]

[16]The full battery of these alternative social-cognitive channels and their corresponding statistical results

Table 10: Narrative-based empathy

	Affective		Cognitive	
	(1)	(2)	(3)	(4)
Treatment	0.024	0.056	-0.032	-0.022
	(0.076)	(0.086)	(0.093)	(0.099)
Constant	5.207***	4.079***	5.313***	5.818***
	(0.066)	(0.751)	(0.042)	(0.797)
Control mean	5.198	5.198	5.306	5.306
Observations	175	175	175	175
R^2	0.119	0.173	0.064	0.085
Wave fixed effects	✓	✓	✓	✓
Controls		✓		✓

Notes: Entries report coefficients from linear regressions of empathy outcomes on an indicator for assignment to the treatment condition. The dependent variables are respondent-level indices constructed from post-reading narrative-response items. *Affective empathy* is the mean of items 1–5, with item 4 reverse-coded, and captures emotional resonance with others' feelings. *Cognitive empathy* is the mean of items 9–14, with items 12 and 13 reverse-coded, and captures perspective-taking and understanding of others' emotions and situations. Higher values indicate greater empathy in each domain. All specifications include first-wave fixed effects. Even-numbered columns additionally include demographic controls. Standard errors clustered at the session level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The complete empathy table, including Sympathy, is reported in Online Appendix Table B.41.

We also administered the revised Reading the Mind in the Eyes Test (RMET) (Baron-Cohen et al., 2001), which measures the ability to infer mental states from the eye region of a face. We find little evidence of improved RMET performance: the controlled treatment estimate is 0.19 additional correct responses and is imprecisely estimated. Thus, the course's effect on narrative engagement does not translate into an improvement in this standard measure of affective theory of mind.

The behavioral effects could also operate through a stronger moral self-concept rather than through a change in how participants perceive others. We examine this possibility using the Moral Identity Scale (Aquino & Reed, 2002), which distinguishes the internalization of moral traits[17] into the self-concept from their public symbolization. We find no evidence of treatment effects on either dimension. Thus, the increase in altruism and trustworthiness is not accompanied by an increase participants' moral self-concept.

Additionally, because the IOS results show lower self–other overlap for some social groups, one possible concern is that the fiction course simply made participants feel more socially

are presented in Online Appendix B.5.9.

[17]Specifically: Caring, Compassionate, Fair, Friendly, Generous, Helpful, Hardworking, Honest, and Kind.

isolated. We examine this possibility using the UCLA Loneliness Scale (Russell, 1996). The 20-item scale has high internal consistency in our sample (Cronbach's $\alpha = 0.93$), with higher values indicating greater perceived loneliness. We find no evidence of a treatment effect on the overall index: the treatment effect is 0.07 points relative to a placebo mean of 2.30. The social group-specific reductions in IOS are therefore not accompanied by a detectable generalized increase in perceived social isolation.

5.4 Robustness

Finally, we perform several robustness checks on our main results. 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. Online Appendix B.5.1 examines this possibility. We find no statistically significant differences by treatment assignment in attendance at at least one course session, completion of at least three sessions, or availability of dictator- or ultimatum-game outcomes. We additionally examine whether participation is differentially associated with observed baseline characteristics across treatment arms by interacting treatment assignment with the full set of baseline covariates, including wave. These interactions are jointly insignificant for all four participation measures. Taken together, these results provide some reassurance against differential selection on observed characteristics, although they do not rule out selection on unobserved determinants of the outcomes.

The estimated treatment effects could partly reflect a detrimental effect of the behavioral-science placebo rather than a positive effect of the fiction course. We compare outcomes within the placebo arm by degree of course exposure, distinguishing participants who attended at least three sessions from those with lower participation (see Online Appendix B.5.5). We find little evidence that greater exposure to the placebo course is associated with lower prosocial behavior or systematically different psychological outcomes. This reduces the concern that the main treatment effects arise because the placebo course itself reduced prosociality

Because the experiment examines several related behavioral and psychological outcomes, we also assess the robustness of the results to multiple-hypothesis testing. We group outcomes into conceptual families corresponding to the main claims of the paper and compute Benjamini–Hochberg adjusted p-values within each family. The core prosocial-behavior family contains dictator-game giving and trust-game trustworthiness; the remaining families cover secondary behavioral outcomes, norm-related outcomes, the main measures of individuation and richer representation, and other psychological outcomes and alternative channels. Full results are reported in Online Appendix B.5.6. The main behavioral pattern is robust to this adjustment, although statistical precision is reduced. In the specifications with controls, the Benjamini–Hochberg adjusted p-values for both dictator-game giving and trustworthiness are 0.084. Thus, both main behavioral effects remain statistically significant at the 10% level after adjusting for multiple testing within the core prosocial-behavior family.

The adjustment also reinforces the distinction between these outcomes and the secondary behavioral outcomes, for which we do not find robust treatment effects after accounting for multiplicity.

We next examine whether the treatment effects persist beyond the session conducted immediately after the course. We compare outcomes measured immediately after the course with those measured at follow-up (see Online Appendix B.5.7). The broad pattern of greater pro-sociality is similar across the two horizons. For the psychological outcomes, the horizon-specific estimates are often imprecise. The estimated effect on Attention to Individual Differences is positive and similar in magnitude across horizons, but the longer-term confidence interval is substantially wider and includes zero. This pattern does not provide clear evidence that the effect diminishes over time.

We also explore whether the treatment effects vary with reading habits measured before random assignment (Online Appendix B.5.8). We summarize the baseline measures into two dimensions: fiction-oriented reading and instrumental reading. The latter primarily captures non-fiction and work- or study-related reading. The results do not reveal a uniform pattern of heterogeneity. There is little evidence that the effect on dictator-game giving varies with fiction-oriented reading, although the estimated effect declines with instrumental reading. By contrast, the effect on trustworthiness is somewhat larger among participants with more fiction-oriented reading habits. Both interactions are significant at the 10% level, but neither is statistically significant when the two reading dimensions and their treatment interactions are included jointly

These exercises reduce several alternative concerns about the main results. The effects on altruism and trustworthiness remain robust to adjustments for multiple-hypothesis testing and are not readily explained by differential attrition, a short-lived post-course response, or a negative effect of greater exposure to the placebo course.

6 Discussion

Our findings provide causal evidence that assignment to a guided literary-fiction course can increase prosocial behavior measured through revealed choices rather than self-reports alone. Relative to a placebo course, participants in the fiction condition give more in the dictator game and behave more trustworthily in the trust game. These treatment effects address the selection concern that people who choose to read fiction may differ systematically from those who do not (Dubourg & Baumard, 2023). At the same time, the treatment does not produce comparable effects on trusting behavior by first movers, ultimatum-game behavior, or third-party punishment. The evidence therefore points to increases in particular prosocial margins—especially altruism and trustworthiness—rather than to a uniform increase in prosocial behavior across strategic and norm-enforcement environments.

The conceptual framework provides one interpretation of this pattern. A finer representation

of others need not increase average other-regarding valuation. Instead, it restores person-specific heterogeneity that a coarse categorical representation compresses: some other-regarding weights may rise and others fall while their mean remains unchanged. When prosocial action begins only once the weight placed on another person is sufficiently high, this spread can nevertheless increase expected prosocial action. The framework therefore gives a precise sense in which fiction can affect behavior by changing how other people are represented without uniformly increasing closeness, similarity, empathy, or average concern for others.

The psychological results line up with this interpretation in several ways. First, the course changed how participants engaged with literary material. Treated participants report greater transportation, cognitive access, narrative presence, and involvement when reading a common story. This suggests that the intervention did not simply expose participants to fiction, but changed the way they read. Because close attention to characters' motives, circumstances, and perspectives was the practice through which the curriculum sought to exercise moral imagination, these results provide evidence that participants engaged in the kind of reading the course was designed to train.

Second, treated participants show greater attention to individual differences on the narrower individuation measure and endorse richer explanations of human behavior on several dimensions of the Attributional Complexity Scale. The broad Uniqueness Orientation index does not change significantly, so the evidence does not support a general shift across every dimension of individuation. Rather, the pattern points more specifically toward greater attention to person-specific differences and to multiple inferred motives and circumstances when explaining behavior.

The IOS results add a distinct piece of evidence. The treatment does not make others uniformly feel closer or more similar to the self. Self-other overlap is largely unchanged for the closest relationships, while it falls for some socially meaningful but less intimate groups. This pattern is consistent with differentiation rather than assimilation: a finer representation may reveal ways in which another person differs from oneself rather than simply making that person feel more like oneself. The absence of robust effects on empathy, RMET, and moral identity further distinguishes the observed pattern from generalized empathy, theory-of-mind, or moral-self-concept accounts.

A further implication of the results is that prosocial behavior can increase without a broad tightening of social norms. We do not observe a systematic pattern of changes in empirical or normative expectations that mirrors the behavioral effects; in some settings, participants instead become more permissive toward deviations from prosocial behavior. This is consistent with a mechanism operating through how individuals represent others and the other-regarding weights entering choice, rather than through an overall shift in beliefs about what others do or what others approve.

More broadly, these findings help clarify what kind of educational intervention a guided fiction-reading course may be. Its value may lie not in directly teaching a prosocial norm,

or in increasing morality or empathy, but in encouraging more differentiated representations of other people. Literary reading may matter not by telling readers what is right, but by changing the social world they perceive when making choices about others—turning, in Brodsky's terms, an interchangeable "zero" into a psychologically distinct person.

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.

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

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

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, *24*(2), 265–272. https://doi.org/10.1214/aoms/1177729032

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, *131*(4), 1753–1794. https://doi.org/10.1093/qje/qjw029

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/].

Bursztyn, L., González, A. L., & 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

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

Cappelen, A., List, J., Samek, A., & 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

Castano, E., Paladino, M. P., Cadwell, O. G., Cuccio, V., & Perconti, P. (2021). Exposure to literary fiction is associated with lower psychological essentialism. *Frontiers in Psychology*, *12*, 662940. https://doi.org/10.3389/fpsyg.2021.662940

Castano, E., Zanella, J., Saedi, F., Zunshine, L., & Ducceschi, L. (2024). On the complexity of literary and popular fiction. *Empirical Studies of the Arts*, *42*(1), 281–300. https://doi.org/10.1177/02762374231163483

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*.

Dodell-Feder, 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

Dovidio, J. F. (1984). Helping behavior and altruism: An empirical and conceptual overview. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 361–427, Vol. 17). Academic Press. https://doi.org/10.1016/S0065-2601(08)60123-9

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

Fryer, J., Roland G., & Jackson, M. O. (2008). A categorical model of cognition and biased decision making. *The B.E. Journal of Theoretical Economics*, *8*(1). https://doi.org/10.2202/1935-1704.1357

Gentzkow, M., & Kamenica, E. (2016). A rothschild-stiglitz approach to bayesian persuasion. *American Economic Review*, *106*(5), 597–601. https://doi.org/10.1257/aer.p20161049

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*.

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–271. https://doi.org/10.1146/annurev.psych.47.1.237

Hoxby, C. M. (2022). The fork in the road: The imperative of investing in adolescent education [Tanner Lectures on Human Values, University of California, Berkeley, April 12–14, 2022]. Retrieved September 6, 2026, from https://tannerlectures.berkeley.edu/caroline-hoxby/

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

Lee, V. K., Kranton, R. E., Conzo, P., & Huettel, S. A. (2021). The hidden cost of humanization: Individuating information reduces prosocial behavior toward in-group members. *Journal of Economic Psychology*, *86*, 102424. https://doi.org/10.1016/j.joep.2021.102424

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1*(3), 593–622. https://doi.org/10.1006/redy.1998.0023

Mullainathan, S., Schwartzstein, J., & Shleifer, A. (2008). Coarse thinking and persuasion. *The Quarterly Journal of Economics*, *123*(2), 577–619. https://doi.org/10.1162/qjec.2008.123.2.577

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

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*.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407. https://doi.org/10.1016/0010-0285(72)90014-X

Rorty, R. (1989). *Contingency, irony, and solidarity*. Cambridge University Press.

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

A.1 Coarse social representations and prosocial choice

This appendix provides the formal version of the argument summarized in Section 2. The purpose is to isolate the consequences of representational granularity. The move from a coarse to a finer representation is taken as exogenous; the experiment provides evidence on whether the fiction intervention is consistent with such a shift.

A.1.1 Social representations

Let Ω be a finite set of inferred social stories, with $\omega \in \Omega$ distributed according to μ. The distribution μ can be interpreted as the decision-maker's subjective distribution over inferred stories. An inferred story summarizes possible person-specific features potentially relevant for social valuation, including inferred motives, circumstances, histories, intentions, or traits. Let

$$v(\omega) \in \mathbb{R}$$

denote the social valuation associated with inferred story ω when the distinctions contained in that story are behaviorally available. The function v is subjective and is left in reduced form. Person-specific features may affect welfare weights through perceived intentions, deservingness, identity, or beliefs about type; type-dependent social-preference models such as Levine (1998) provide one possible foundation.

A social representation is a partition

$$\Pi = \{P_1, \ldots, P_K\}$$

of Ω. For a story ω, let $\Pi(\omega)$ denote the cell containing it. Stories in the same cell are treated as interchangeable for purposes of social valuation. Their common other-regarding weight is

$$\beta_\Pi(\omega) \equiv \mathbb{E}_\mu[v(\tilde{\omega}) \mid \tilde{\omega} \in \Pi(\omega)] .$$

This conditional-mean rule need not describe literal mental averaging. If the decision-maker cannot distinguish the stories pooled within a cell and social valuation enters utility linearly, the unresolved stories affect expected utility only through their conditional mean. The rule is also compatible with prototype accounts in which heterogeneous cues are compressed into an abstracted category representation (Reed, 1972). We thereby isolate coarseness from additional within-category distortions such as representativeness or salience (Bordalo et al., 2016).

Consider two representations, Π_C and Π_F, where Π_F refines Π_C: every cell of Π_F is contained in a cell of Π_C. The refinement leaves μ and $v(\omega)$ unchanged and changes only representational granularity. Let

$$\beta_C \equiv \beta_{\Pi_C}, \qquad \beta_F \equiv \beta_{\Pi_F}.$$

Proposition 1. *If Π_F refines Π_C, then*

$$\beta_C = \mathbb{E}[\beta_F \mid \Pi_C].$$

Consequently,

$$\mathbb{E}[\beta_F] = \mathbb{E}[\beta_C] = \mathbb{E}[v(\omega)]$$

and

$$\beta_C \preceq_{cx} \beta_F,$$

where $\beta_C \preceq_{cx} \beta_F$ means that $\mathbb{E}[\phi(\beta_C)] \leq \mathbb{E}[\phi(\beta_Y)]$ for every convex function ϕ for which the expectations exist. Equivalently, β_F is a mean-preserving spread of β_C. Moreover,

$$\mathrm{Var}(\beta_F) = \mathrm{Var}(\beta_C) + \mathbb{E}[\mathrm{Var}(\beta_F \mid \Pi_C)] \geq \mathrm{Var}(\beta_C).$$

With strict inequality whenever $\Pr(\mathrm{Var}(\beta_F \mid \Pi_C) > 0) > 0$.

Proof. Because Π_F refines Π_C, the finer cells contained in any coarse cell partition that coarse cell. Averaging their other-regarding weights therefore recovers the average valuation assigned to the coarse cell, which gives the martingale property, $\beta_C = \mathbb{E}[\beta_F \mid \Pi_C]$. The law of iterated expectations gives equality of means.

For any convex function ϕ, conditional Jensen's inequality gives

$$\phi(\beta_C) = \phi(\mathbb{E}[\beta_F \mid \Pi_C]) \leq \mathbb{E}[\phi(\beta_F) \mid \Pi_C].$$

Taking expectations yields $\mathbb{E}[\phi(\beta_C)] \leq \mathbb{E}[\phi(\beta_F)]$, establishing β_C is dominated by β_F in the convex order. Finally, the law of total variance applied to β_F conditional on Π_C, together with the martingale property, yields the larger variance for the finer representation. The variance inequality is strict whenever the finer representation reveals valuation-relevant differences within at least one coarse category with positive probability. □

Proposition 1 is an application of standard conditional-expectation and information-ordering logic (Blackwell, 1953; Gentzkow & Kamenica, 2016). In this application it means that individuation "decompresses" social valuation. Under coarse reasoning, distinct stories receive the average valuation of their category; under a finer representation, some can receive higher social weights and others lower social weights while average valuation remains unchanged. A finer partition has no effect when the distinctions it reveals are irrelevant for $v(\omega)$.

A.1.2 Prosocial choice

Suppose the decision-maker chooses a prosocial action $a \geq 0$ that produces a units of benefit for the other person that the decision-maker values according to β. Private cost is $\kappa a + \frac{1}{2}a^2$, with $\kappa \geq 0$, so utility is

$$u(a; \beta) = \beta a - \kappa a - \frac{1}{2}a^2.$$

The parameter κ is the marginal private cost of prosocial action at zero. Positive action therefore requires the social value of benefiting the other to be sufficiently large relative to its private cost. This reduced-form trade-off is consistent with classic cost–reward accounts of helping (Dovidio, 1984).[18]

Maximizing $u(a;\beta)$ over $a \geq 0$ gives

$$a^*(\beta) = (\beta - \kappa)_+ \equiv \max\{\beta - \kappa, 0\}.$$

The response is convex in β: below κ, social concern does not translate into positive action; above the threshold, additional social weight increases action one for one.

Proposition 2. *If Π_F refines Π_C, then under the choice problem with $u(a;\beta)$,*

$$\mathbb{E}[a^*(\beta_F)] \geq \mathbb{E}[a^*(\beta_C)]. \tag{4}$$

The inequality is strict if there exists a coarse category with positive probability for which the finer representation generates other-regarding weights both strictly below and strictly above κ.

Proof. By Proposition 1, $\beta_C \preceq_{cx} \beta_F$. Since $\phi_\kappa(\beta) = (\beta - \kappa)_+$ is convex, the definition of convex order implies

$$\mathbb{E}[(\beta_C - \kappa)_+] \leq \mathbb{E}[(\beta_F - \kappa)_+],$$

which is (4). Strict inequality requires conditional Jensen's inequality to be strict in at least one coarse category. Because ϕ_κ is linear on either side of the kink at κ, this occurs when the refinement reveals other-regarding weights on both sides of the threshold with positive probability. □

A.1.3 Illustration and interpretation

Consider two equally likely stories with $v(\omega_1) = 0$ and $v(\omega_2) = 2$. If the coarse representation pools them, both receive $\beta_C = 1$. If the finer representation distinguishes them, the other-regarding weights are $\beta_F \in \{0, 2\}$ with equal probability. The mean social weight remains one. Let $\kappa = 1$. Under the coarse representation $a^*(1) = 0$, whereas under the fine representation $a^*(0) = 0$ and $a^*(2) = 1$, so

$$\mathbb{E}[a^*(\beta_F)] = \frac{1}{2} > 0 = \mathbb{E}[a^*(\beta_C)].$$

Although average social weight is unchanged, refinement recovers distinctions that were previously averaged away. Because prosocial action responds convexly to social weight, these distinctions can increase expected prosocial action.

[18] In allocation settings such as dictator-game transfers or trust-game returns, this private cost also has a literal opportunity-cost interpretation: increasing the other's payoff requires sacrificing one's own payoff.

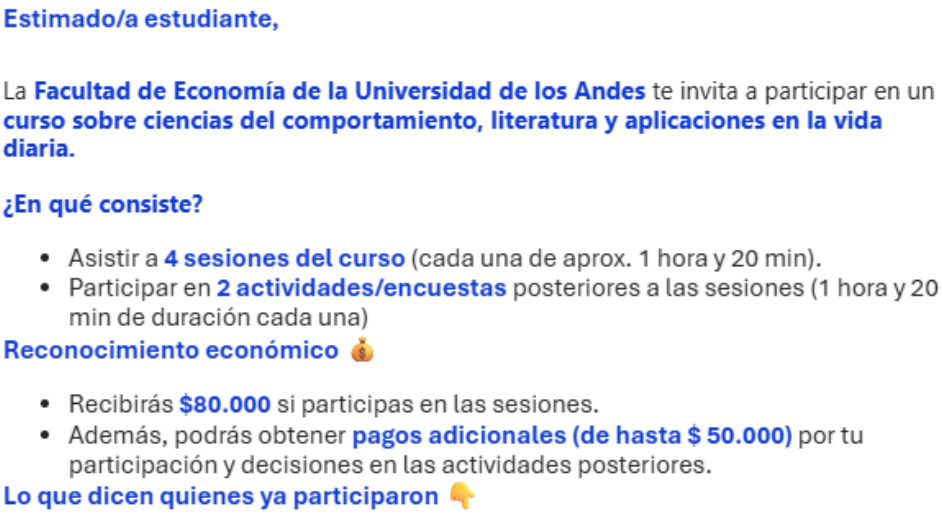
The qualification is important. Proposition 1 is a result about representation: refinement spreads social weights. Proposition 2 additionally requires a nonlinear behavioral response. If refinement reveals variation only in a region where action is linear, expected action need not change; if the relevant choice is bounded above, the optimal response need not be globally convex. The model therefore does not imply that finer representation always increases prosociality, nor does it imply that knowing more about another person uniformly raises the weight placed on that person's welfare.

B Online Appendix

B.1 Recruitment

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



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 B.1: Invitation email

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

B.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.

B.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.

B.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.

Table B.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.

Perceived Uniqueness / Individuation (study-developed scale) Table B.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.

Table B.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).

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 B.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 B.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?	

#	Item	R?
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 B.5). The instrument draws on the Transportation framework of Green and Brock (2000b) and the multidimensional Narrative Engagement 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 B.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**
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.	

#	Item (English translation)	R?
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**
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 B.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 B.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 B.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.	

#	Item	R?
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 B.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 B.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.	

#	Item	R?
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**
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."	–

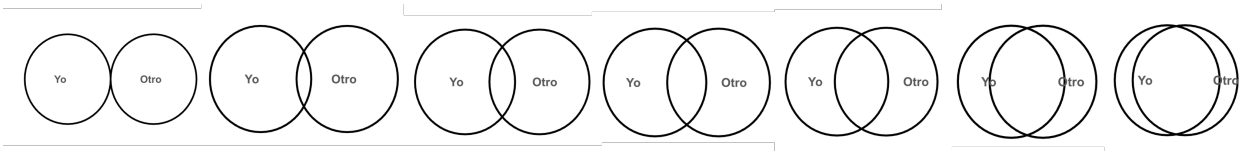


Figure B.4: Inclusion of the Other in the Self pictorial options

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 B.4

Table B.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

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 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.

B.5 Additional results and robustness checks

B.5.1 Selection into the Analytic Sample

Table B.9: Participation and Outcome Availability

	Attended ≥1 session		Attended ≥3 sessions		Set A observed		Set B observed	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Baseline	With controls	Baseline	With controls	Baseline	With controls	Baseline	With controls
Assigned treatment	−0.046	−0.047	0.031	0.034	0.011	0.010	0.036	0.034
	(0.039)	(0.038)	(0.040)	(0.039)	(0.036)	(0.035)	(0.036)	(0.036)
First wave		−0.108		−0.112		0.110		0.078
		(0.049)		(0.050)		(0.045)		(0.046)
Age		0.004		−0.015		0.019		−0.013
		(0.009)		(0.010)		(0.009)		(0.009)
Quiz score		0.0004		−0.0004		0.0024		0.0001
		(0.0017)		(0.0017)		(0.0016)		(0.0016)
Knows others		0.036		0.193		0.122		0.017
		(0.068)		(0.070)		(0.062)		(0.063)
Female		0.020		0.013		−0.007		0.019
		(0.039)		(0.040)		(0.036)		(0.036)
High strata		0.275		0.231		0.104		0.217
		(0.063)		(0.064)		(0.058)		(0.058)
Economics field		−0.030		−0.017		0.002		−0.059
		(0.051)		(0.053)		(0.047)		(0.048)
Constant	0.657	0.554	0.422	0.694	0.256	−0.212	0.253	0.476
	(0.028)	(0.198)	(0.029)	(0.203)	(0.026)	(0.182)	(0.026)	(0.184)
Placebo mean	0.657	0.657	0.422	0.422	0.256	0.256	0.253	0.253
Observations	618	618	618	618	618	618	618	618
Observations with outcome = 1	391	391	271	271	162	162	168	168
R^2	0.002	0.044	0.001	0.050	0.0002	0.027	0.002	0.034
Joint treatment-interaction p-value		0.430		0.146		0.275		0.132

Notes: All regressions use the same full analysis sample of 618 invited students: 289 assigned to placebo and 329 assigned to treatment. The dependent variables are indicators for attending at least one course session (columns 1–2), attending at least three course sessions (columns 3–4), having Set A data observed, dictator-game included (columns 5–6) and having Set B data observed, ultimatum-game included (columns 7–8) as described in Section 3. Odd-numbered columns include treatment assignment and an intercept. Even-numbered columns additionally include age, baseline quiz score, a first-wave indicator, whether the participant knew other subjects, female, high socioeconomic stratum, and economics field. "Placebo mean" reports the dependent-variable mean among observations assigned to placebo. "Observations with outcome = 1" reports the number satisfying the corresponding participation or availability criterion across both treatment arms. The final row reports the p-value from a separate regression that adds interactions between treatment assignment and all seven listed covariates, including wave. The null hypothesis is that all seven interaction coefficients jointly equal zero. The coefficients displayed above come from the specifications without these interactions. Robust standard errors are reported in parentheses. * p<0.1, ** p<0.05, *** p<0.01.

Table B.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)	4.11	4.14	4.44	3.74	0.18
Fiction books read (# of books)	3.16	4.05	3.29	2.89	0.45
Non-fiction books read (# of books)	1.90	2.60	2.11	1.67	0.20
Work- or study-related books read (# of books)	5.46	7.97	5.36	5.56	0.86
Fiction books purchased (# of books)	1.68	3.26	1.60	1.76	0.71
Non-fiction books purchased (# of books)	1.05	2.14	1.25	0.84	0.15
Work- or study-related books purchased (# of books)	0.92	1.96	0.96	0.89	0.80
Fiction increases empathy (likert 1–4)	2.88	0.67	2.87	2.90	0.70
Fiction increases intelligence (likert 1–4)	2.73	0.59	2.73	2.71	0.80
Fiction helps understand complex motivations (likert 1–4)	2.99	0.72	3.02	2.96	0.51
Fiction helps recognize individual uniqueness (likert 1–4)	2.85	0.72	2.81	2.90	0.35
Observations			133	114	

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.

B.5.2 Trust Game

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

	Personal norm		Normative expectations	
	(1)	(2)	(3)	(4)
Treatment	-0.094	-0.125	-0.006	-0.026
	(0.126)	(0.138)	(0.051)	(0.047)
Constant	2.538***	2.318***	2.330***	2.114***
	(0.121)	(0.549)	(0.033)	(0.391)
Placebo mean	2.538	2.538	2.330	2.330
Observations	190	190	190	190
R^2	0.003	0.053	0.000	0.033
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 B.12: Player 2 (Trustworthiness) Behavior and Empirical Expectations

	% sent back (P2)		expected % sent back	
	(1)	(2)	(3)	(4)
	Baseline	With controls	Baseline	With controls
Treatment	0.080***	0.079***	0.095*	0.106**
	(0.030)	(0.028)	(0.044)	(0.039)
Amount sent (P1)	0.043***	0.043***	-0.010**	-0.010**
	(0.003)	(0.003)	(0.004)	(0.003)
Treatment × Amount sent (P1)	-0.007*	-0.007*	-0.011	-0.013*
	(0.004)	(0.004)	(0.007)	(0.006)
Constant	0.058***	0.045**	0.360***	0.387***
	(0.017)	(0.022)	(0.026)	(0.028)
Placebo mean	0.296	0.296	0.294	0.294
Observations	1900	1900	186	186
R^2	0.388	0.411	0.145	0.182
Wave fixed effects	✓	✓	✓	✓
Demographic controls		✓		✓

Standard errors clustered at the individual level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
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.

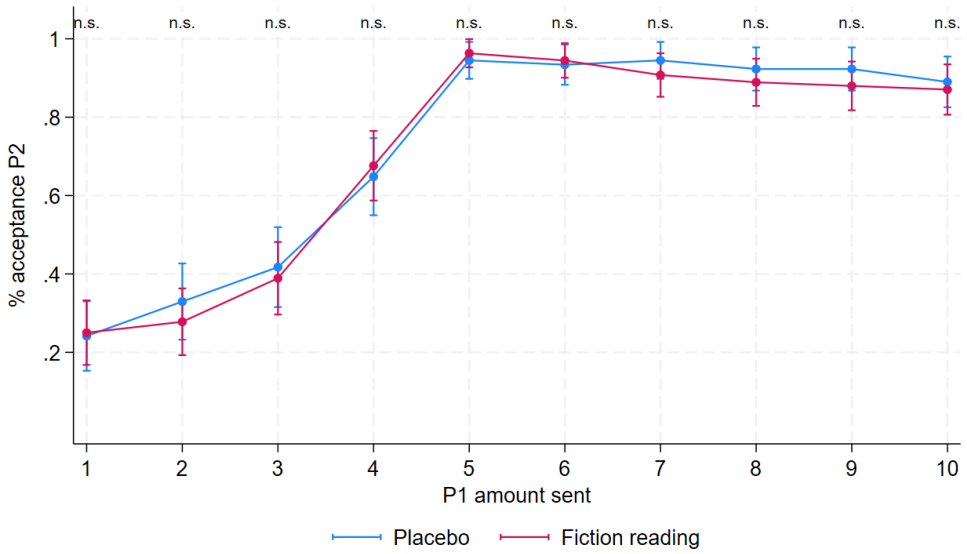
B.5.3 Ultimatum Game

Table B.13: Player 1 (Ultimatum) Behavior and Empirical Expectations

	Amount sent (P1)		P2 expectations about P1 sent	
	(1)	(2)	(3)	(4)
Treatment	0.233	0.211	0.049	0.039
	(0.136)	(0.126)	(0.134)	(0.150)
Constant	4.672***	4.625***	4.385***	4.285***
	(0.094)	(0.507)	(0.107)	(0.519)
Placebo mean	4.670	4.670	4.385	4.385
Observations	199	199	199	199
R^2	0.016	0.059	0.001	0.045
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 B.5: 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 B.14: 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	-0.000	0.001	-0.008	-0.007
	(0.037)	(0.038)	(0.052)	(0.051)
Constant	0.338***	0.533***	0.416***	0.540**
	(0.026)	(0.139)	(0.038)	(0.213)
Control mean	0.338	0.338	0.418	0.418
Observations	199	199	199	199
R^2	0.012	0.059	0.020	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 B.15: Responder Acceptance Decisions

	(1)	(2)
Treatment	0.012	0.007
	(0.053)	(0.053)
P1 Offer	0.091***	0.091***
	(0.006)	(0.006)
Treatment × P1 Offer	-0.004	-0.004
	(0.008)	(0.008)
Constant	0.204***	0.206***
	(0.038)	(0.038)
Placebo mean	0.659	0.659
Observations	2189	2189
R^2	0.35	0.35
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

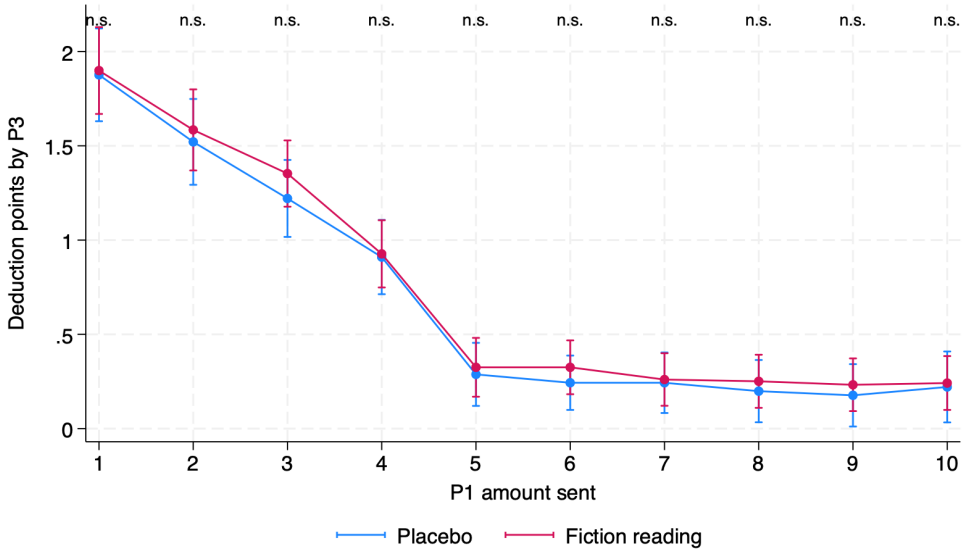
B.5.4 Third-party Game

Table B.16: Player 1 (Third party) Behavior and Empirical Expectations

	Amount sent		Expectations	
	(1)	(2)	(3)	(4)
Treatment	0.045	-0.008	-0.179	-0.218
	(0.131)	(0.119)	(0.160)	(0.188)
Constant	4.624***	2.624*	4.394***	3.727***
	(0.099)	(1.281)	(0.048)	(0.558)
Placebo mean	4.626	4.626	4.396	4.396
Observations	199	199	199	199
R^2	0.008	0.056	0.007	0.025
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 B.6: 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 B.17: 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	0.041	0.041	0.060	0.059
	(0.183)	(0.182)	(0.176)	(0.176)
Amount sent (P1)	-0.204***	-0.204***	-0.207***	-0.207***
	(0.018)	(0.018)	(0.019)	(0.019)
Treatment × Amount sent (P1)	0.002	0.002	-0.002	-0.002
	(0.025)	(0.025)	(0.026)	(0.026)
Placebo mean	0.775	0.775	0.817	0.817
Observations	2186	2186	2189	2189
R^2	0.296	0.305	0.310	0.314
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 B.18: Player 2 (Third party) Personal and social Norms regarding Player 1

	Personal norm		Social norm	
	(1)	(2)	(3)	(4)
Treatment	0.021	0.039	-0.113	-0.104
	(0.086)	(0.080)	(0.130)	(0.130)
Constant	1.953***	1.965***	1.810***	1.987***
	(0.064)	(0.261)	(0.122)	(0.415)
Placebo mean	1.945	1.945	1.802	1.802
Observations	199	199	199	199
R^2	0.248	0.281	0.267	0.289
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 B.19: Player 2 (Third party) Personal and social Norms regarding Player 3

	Pnorm deductions of P3		Snorm deductions of P3	
	(1)	(2)	(3)	(4)
Treatment	-0.204*	-0.177	-0.082	-0.082
	(0.095)	(0.099)	(0.081)	(0.067)
Constant	3.010***	4.195***	2.974***	3.935***
	(0.064)	(0.241)	(0.067)	(0.245)
Placebo mean	3.000	3.000	2.967	2.967
Observations	199	199	199	199
R^2	0.297	0.357	0.171	0.211
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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.5.5 Placebo and treated group behavior

To examine whether simple exposure to the experimental environment affected behavior in the games, Tables B.20–B.21 and Figures B.7–B.8 analyze the relationship between placebo-session attendance and behavioral outcomes. We first focus on participants assigned to the placebo arm, additionally including individuals who only participated in the experimental games but did not attend any placebo sessions.

Table B.20 compares placebo participants who attended at least three sessions with those who attended fewer than three sessions for Player 1 or proposer behavior. Table B.21 reports the analogous comparison for other-player behavior. Across specifications, the results do not reveal a consistent relationship between high placebo exposure and experimental behavior. Some coefficients are statistically significant in isolated specifications, particularly for selected outcomes, but the direction, magnitude, and robustness of these associations are not stable across games, player roles, or model parameterizations.

The number of observations differs across outcomes because the realized data do not contain a complete game battery for every participant. For example, in the Trust game, 103 placebo participants have Player 1 behavior observed, while 84 have the Player 2 conditional return decisions needed to construct Trustworthiness P2. This difference is driven by 19 placebo observations with observed Trust P1 behavior but missing Trust P2 conditional return decisions. It is not driven by only-games participants: 11 only-games participants have Trust P1 observed, and the same 11 also have Trustworthiness P2 observed.

Figures B.7 and B.8 further illustrate the absence of a clear dose-response pattern. Mean behavior does not display a systematic monotonic relationship with the number of placebo sessions attended, and confidence intervals overlap substantially across exposure levels. Individuals who only completed the experimental games, corresponding to zero placebo sessions, generally behave similarly to participants with partial or full placebo exposure.

As an additional descriptive benchmark, Tables B.22 and B.23 compare high-exposure treated participants to pure controls who only completed the experimental games and did not attend any course session. This exercise is informative as a robustness check, although the number of pure controls is small and the estimates should be interpreted with caution. Specifications with controls are particularly sensitive because baseline covariates are missing for many pure controls and must be handled using missing-control indicators.

Overall, these findings suggest that mere exposure to the experimental setting is unlikely to account for the main treatment effects reported in the paper. At the same time, because placebo-session attendance and pure-control status were not randomly assigned, these exercises should be interpreted as descriptive robustness checks rather than causal estimates.

Table B.20: Placebo group results on games: Player 1 behavior

	Dictator P1		Trust P1		Ultimatum P1		TPG P1	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Completed at least 3 sessions	-0.130	-0.378	1.172*	1.591***	-0.067	-0.149	-0.257	-0.392**
	(0.464)	(0.581)	(0.662)	(0.507)	(0.196)	(0.145)	(0.161)	(0.176)
Constant	3.737***	6.468***	5.316***	10.616***	4.714***	5.534**	4.857***	5.688***
	(0.426)	(1.525)	(0.571)	(1.331)	(0.189)	(2.048)	(0.151)	(0.818)
Mean dep. var. (<3 sessions)	3.737	3.737	5.316	5.316	4.714	4.714	4.857	4.857
Observations	103	103	103	103	99	99	99	99
R^2	0.001	0.101	0.027	0.203	0.000	0.102	0.006	0.085
Controls		✓		✓		✓		✓
Clustered SE	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table reports associations between high placebo exposure and Player 1 or proposer behavior among individuals assigned to the placebo arm, including participants that only completed the experimental games. High placebo exposure is defined as completing at least three placebo sessions. The omitted category is participants assigned to placebo who attended fewer than three sessions. Some participants completed different games across the short-term and long-term measurement periods, and the realized data do not contain a complete game battery for every participant; therefore, the number of observations varies across outcomes. Specifications with controls are estimated on the subset of participants with non-missing baseline covariates. Because completion of the placebo course was not randomly assigned, coefficients should be interpreted as correlational associations rather than causal effects. Standard errors clustered at the session-term level are reported in parentheses; participants without a session-term identifier are assigned individual clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.21: Placebo group results on games: Other-player behavior

	Trustworthiness P2		Ultimatum P2		TPG P3	
	(1)	(2)	(3)	(4)	(5)	(6)
Completed at least 3 sessions	-0.077	-0.053	0.206***	0.080	-0.210	-0.313
	(0.052)	(0.053)	(0.062)	(0.047)	(0.169)	(0.233)
Amount sent by P1	0.038***	0.038***				
	(0.006)	(0.006)				
Completed at least 3 sessions × Amount sent by P1	0.007	0.007				
	(0.007)	(0.007)				
Amount sent by P1 / received by P3					-0.203***	-0.203***
					(0.017)	(0.017)
Constant	0.121**	0.269*	0.143***	0.148	1.978***	0.837
	(0.048)	(0.147)	(0.044)	(0.323)	(0.192)	(0.694)
Mean dep. var. (<3 sessions)	0.329	0.329	0.143	0.143	0.961	0.961
Individuals	84	84	99	99	99	99
Observations	840	840	99	99	1087	1087
R^2	0.454	0.485	0.055	0.188	0.297	0.316
Controls		✓		✓		✓
Clustered SE	✓	✓	✓	✓	✓	✓

Notes: This table reports associations between high placebo exposure and other-player behavior among individuals assigned to the placebo arm, including participants that only completed the experimental games. High placebo exposure is defined as completing at least three placebo sessions. The omitted category is participants assigned to placebo who attended fewer than three sessions. Columns (1)–(2) report Player 2 behavior in the Trust game, columns (3)–(4) report Player 2 behavior in the Ultimatum game, and columns (5)–(6) report Player 3 behavior in the Third-party punishment game. Trustworthiness P2 is defined as the percentage of the amount received that Player 2 returns. Ultimatum P2 is the acceptance rate of low offers, defined as offers below the equal split. TPG P3 is the deduction chosen by Player 3; regressions for TPG P3 control for the amount sent by Player 1 / received by Player 3. For Trustworthiness P2 and TPG P3, observations are conditional decisions; for Ultimatum P2, observations are individual-level acceptance rates. The row "Individuals" reports the number of unique experimental observations with the corresponding outcome observed in the estimation sample. Some participants completed different games across the short-term and long-term measurement periods, and the realized data do not contain a complete game battery for every participant; therefore, the number of observations varies across outcomes. Specifications with controls are estimated on the subset of participants with non-missing baseline covariates. Because completion of the placebo course was not randomly assigned, coefficients should be interpreted as correlational associations rather than causal effects. Standard errors clustered at the relevant experimental-observation or session-term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

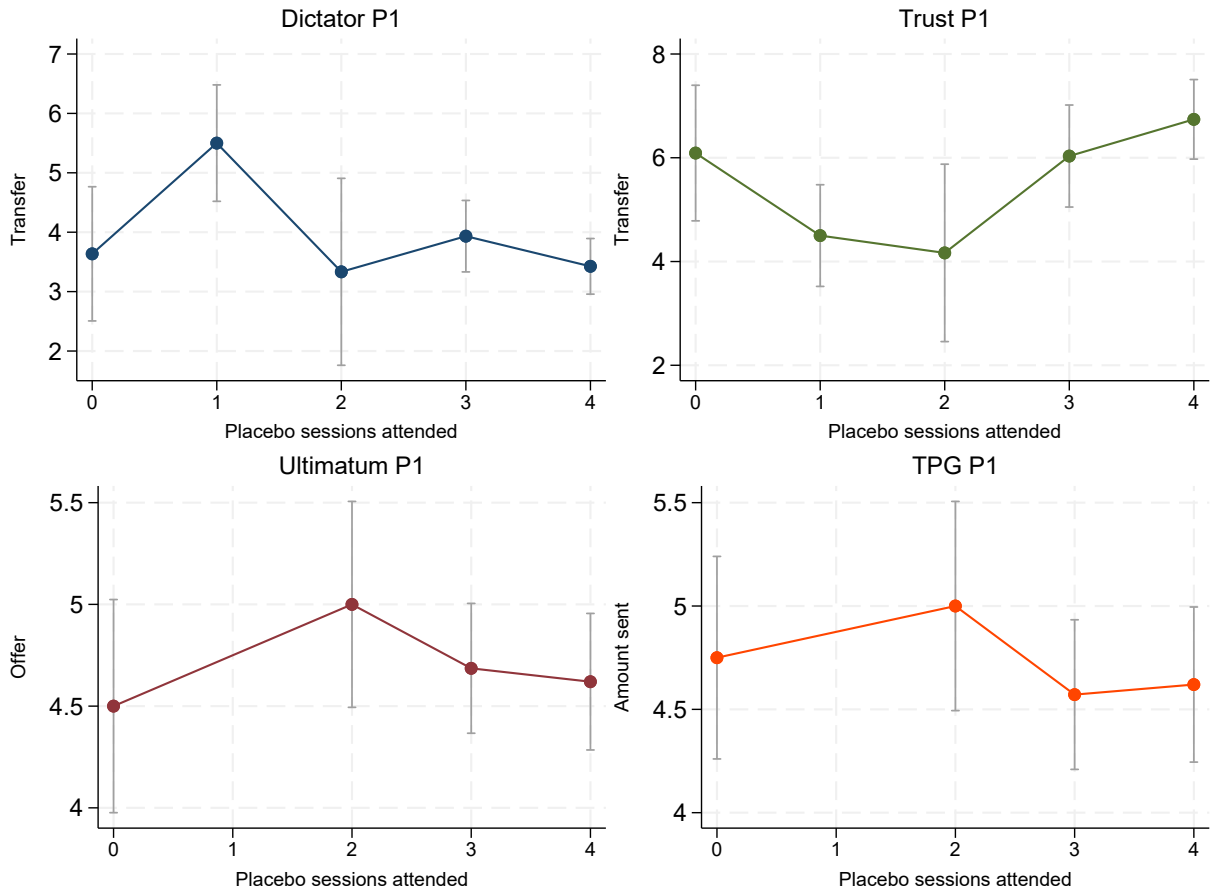


Figure B.7: Placebo-session attendance and Player 1 behavior

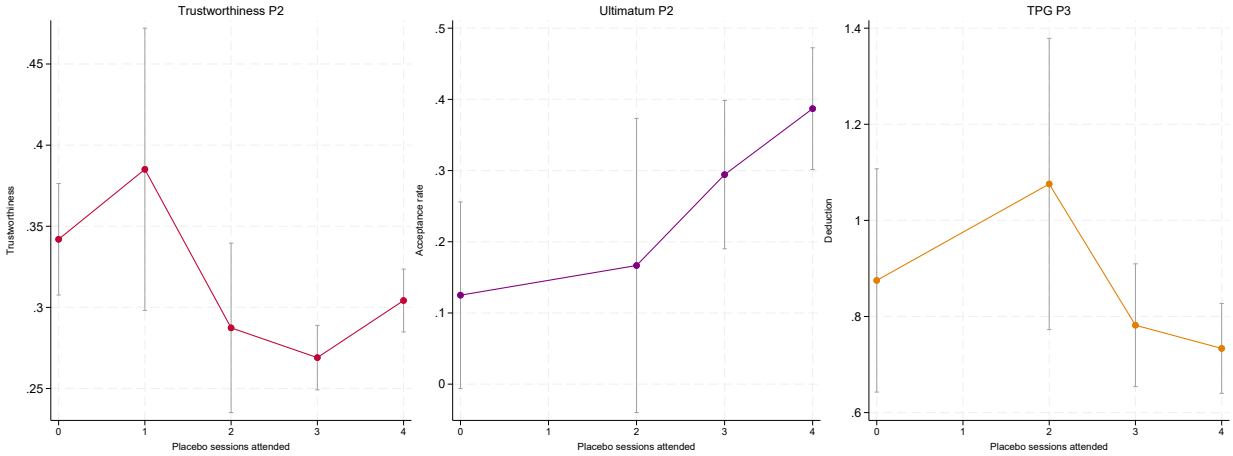


Figure B.8: Placebo-session attendance and other-player behavior

Table B.22: Placebo high-exposure participants versus pure controls: Player 1 behavior

	Dictator		Trust		Ultimatum		Third party	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Placebo, 3+ sessions	-0.029	0.291	0.397	0.644	0.147	0.329	-0.150	-0.028
	(0.585)	(0.594)	(0.748)	(0.791)	(0.280)	(0.374)	(0.254)	(0.356)
Constant	3.636***	6.432***	6.091***	10.992***	4.500***	5.364**	4.750***	5.494***
	(0.569)	(1.737)	(0.658)	(1.199)	(0.260)	(1.952)	(0.243)	(1.067)
Mean dep. var. pure control	3.636	3.636	6.091	6.091	4.500	4.500	4.750	4.750
Pure controls	11	11	11	11	8	8	8	8
Observations	95	95	95	95	93	93	93	93
R^2	0.000	0.089	0.002	0.142	0.001	0.108	0.001	0.094
Controls		✓		✓		✓		✓
Session cluster	✓	✓	✓	✓	✓	✓	✓	✓

Notes: The sample is restricted to placebo participants who attended at least three sessions and pure controls who attended zero sessions. Pure controls are individuals who only completed the experimental games and did not attend any course session. Columns (1)–(2) report results for the Dictator game, columns (3)–(4) for the Trust game, columns (5)–(6) for the Ultimatum game, and columns (7)–(8) for the Third-party punishment game. Each column reports coefficients from a separate linear regression comparing treated high-exposure participants to pure controls. The row "Mean dep. var. pure control" reports the mean of the dependent variable among pure controls in the corresponding estimation sample. Specifications marked with a ✓ in the "Controls" row include baseline covariates and missing-control indicators to preserve the pure-control comparison group. Because attendance and pure-control status were not randomly assigned, estimates should be interpreted as descriptive robustness checks rather than causal effects. The number of pure controls is small, so controlled specifications should be interpreted with caution. Standard errors clustered at the session level are reported in parentheses; pure controls without a session identifier are assigned individual clusters. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.23: Placebo high-exposure participants versus pure controls: Other-player behavior

	Trustworthiness P2		Ultimatum P2		TPG P3	
	(1)	(2)	(3)	(4)	(5)	(6)
Placebo, 3+ sessions	-0.105	-0.096	0.224***	0.196**	-0.123	-0.091
	(0.073)	(0.075)	(0.073)	(0.085)	(0.188)	(0.211)
Amount sent by P1	0.035***	0.035***				
	(0.008)	(0.008)				
Placebo, 3+ sessions × Amount sent by P1	0.009	0.009				
	(0.009)	(0.009)				
Amount sent by P1 / received by P3					-0.199***	-0.199***
					(0.018)	(0.018)
Constant	0.149**	0.330**	0.125*	-0.013	1.868***	0.650
	(0.071)	(0.156)	(0.065)	(0.282)	(0.215)	(0.697)
Mean dep. var. pure control	0.342	0.342	0.125	0.125	0.875	0.875
Pure controls	11	11	8	8	8	8
Individuals	76	76	93	93	93	93
Observations	760	760	93	93	1021	1021
R^2	0.471	0.494	0.042	0.181	0.284	0.299
Controls		✓		✓		✓
Clustered SE	✓	✓	✓	✓	✓	✓

Notes: The sample is restricted to placebo participants who attended at least three sessions and pure controls who attended zero sessions. Pure controls are individuals who only completed the experimental games and did not attend any course session. Column (1) reports Player 2 behavior in the Trust game, column (2) reports Player 2 behavior in the Ultimatum game, and column (3) reports Player 3 behavior in the Third-party punishment game. Trustworthiness P2 is defined as the percentage of the amount received that Player 2 returns. Ultimatum P2 is the acceptance rate of low offers, defined as offers below the equal split. TPG P3 is the deduction chosen by Player 3; the regression controls for the amount sent by Player 1 / received by Player 3. For Trustworthiness P2 and TPG P3, observations are conditional decisions; for Ultimatum P2, observations are individual-level acceptance rates. The row "Individuals" reports the number of unique experimental observations with the corresponding outcome observed in the estimation sample. Because attendance and pure-control status were not randomly assigned, estimates should be interpreted as descriptive robustness checks rather than causal effects. The number of pure controls is small, so estimates should be interpreted with caution. Standard errors clustered at the relevant experimental-observation or session-term level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B.5.6 Multiple hypothesis testing

In this appendix, we report a complementary robustness exercises for the main experimental outcomes: multiple-hypothesis adjustment. 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 two p-values: the conventional p-value from the main specification in parentheses and the Benjamini–Hochberg multiple-hypothesis-adjusted p-value in braces.

Table B.24: MHT on family 1

	Dictator giving		Trustworthiness	
	(1)	(2)	(3)	(4)
Treatment	(0.030)	(0.049)	(0.047)	(0.084)
	{0.047}	{0.084}	{0.047}	{0.084}
Observations	175	171	1380	1350
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 percentage of the amount received that Player 2 returns 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 dictator giving, the reported p-values correspond to the treatment indicator. In trustworthiness, the reported p-values correspond to the treatment indicator from the main Player 2 specification for percentage sent back. 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 B.25: 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}
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 B.26: 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}
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 B.27: 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}
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 B.28: 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}
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 B.29: 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}
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 B.30: MHT on Atributtion scales

	Attribution index		Motivation		Complex		Metacognitive		Behavior		Ambiguity		Situational		Cognitive effort	
	(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}
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 B.31: 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}
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 B.32: 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}
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 B.33: MHT on Loneliness scale

	UCLA loneliness	
	(1)	(2)
Treatment	(0.513)	(0.615)
	{0.513}	{0.615}
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.

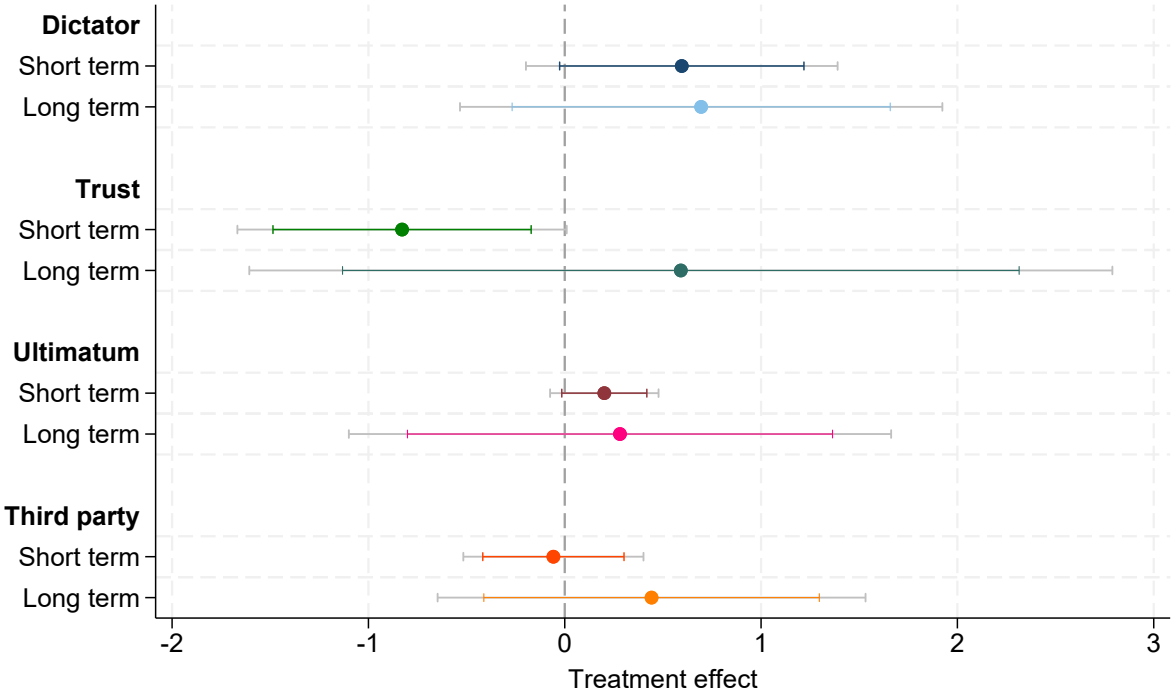
Table B.34: MHT on RMET scale

	RMET	
	(1)	(2)
Treatment	(0.442)	(0.357)
	{0.442}	{0.357}
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.

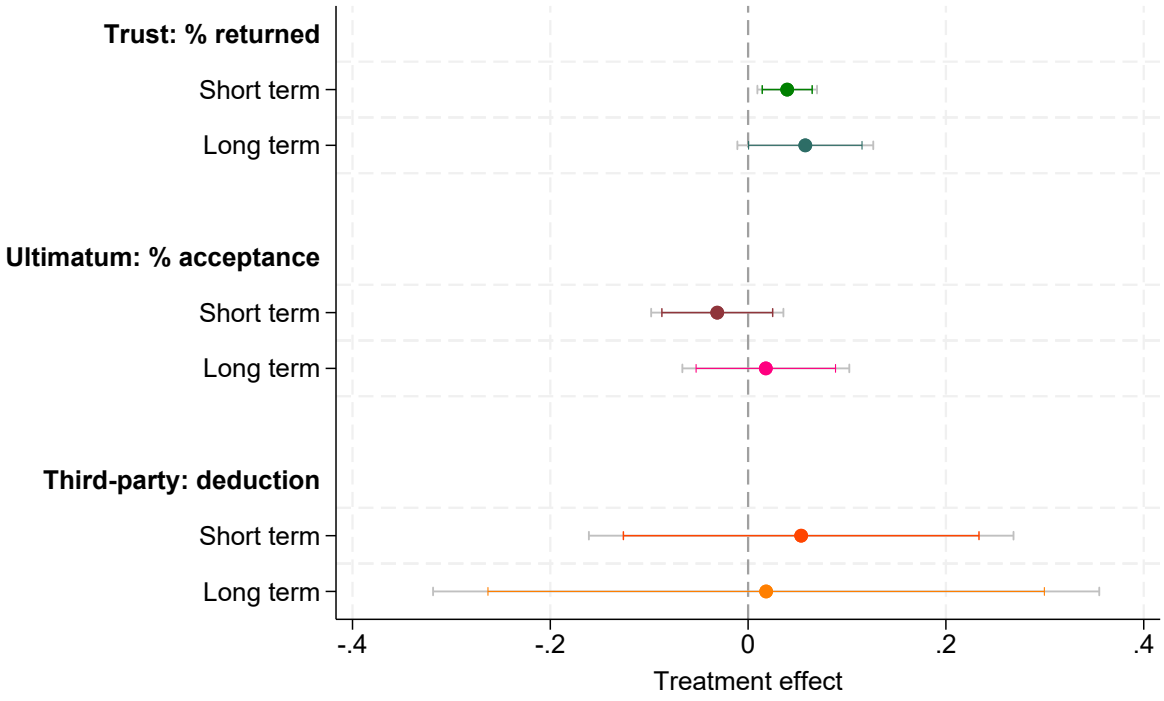
B.5.7 Short vs. Longer-term treatment effects

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



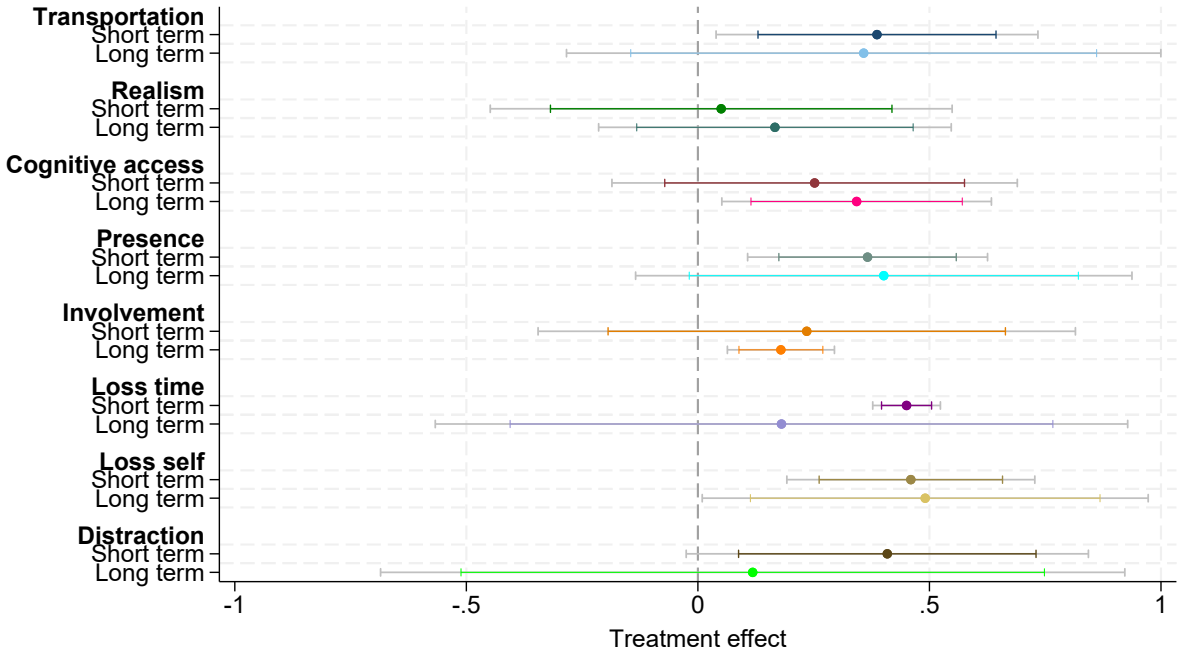
Notes: The figure reports treatment-effect estimates from separate linear regressions of behavioral outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for participants surveyed in the short term and in the long term after treatment assignment. All specifications 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 B.10: Treatment effects of behavior in games (Short term vs Long term)



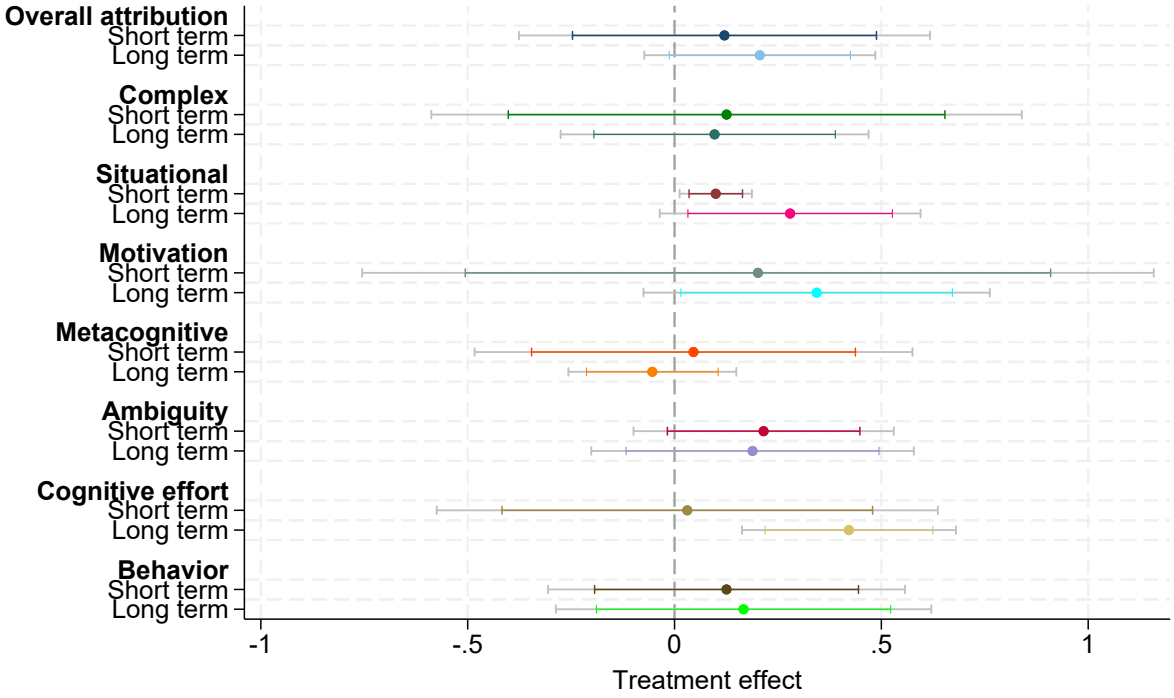
Notes: The figure reports treatment-effect estimates from separate linear regressions of panel outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and longer-term samples. In the trust game, the outcome is the share of the received amount 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 B.11: Treatment effects on Narrative subscales (Short term vs Long term)



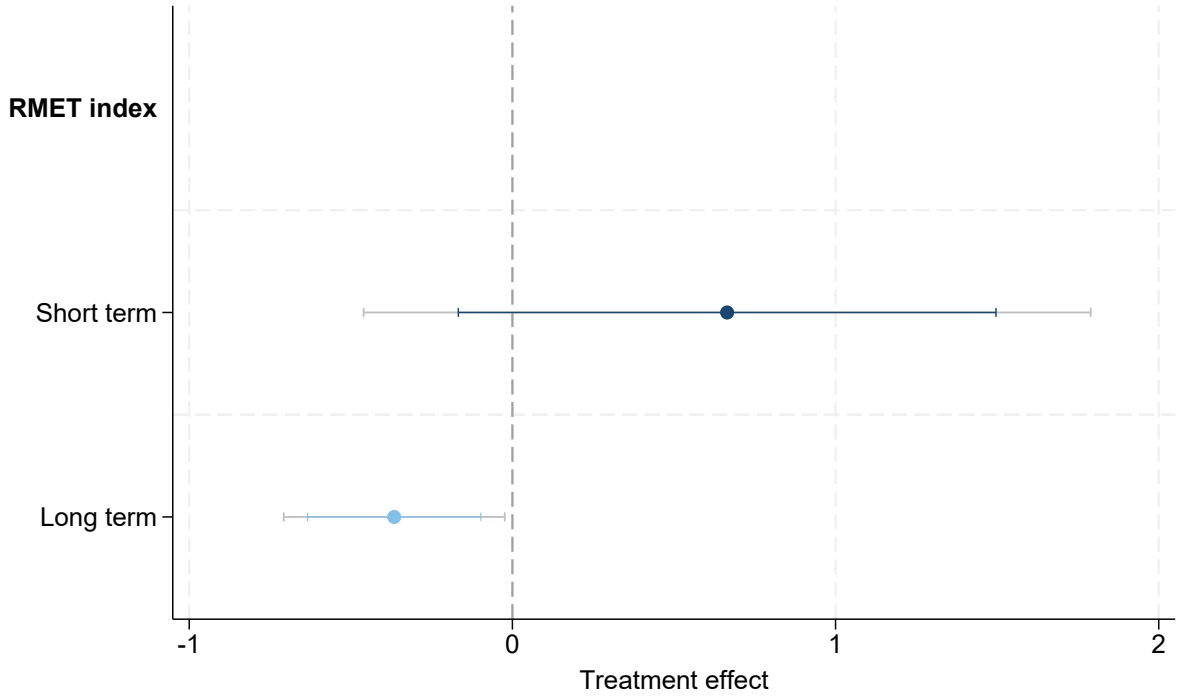
Notes: The figure reports treatment-effect estimates from separate baseline regressions of narrative-scale outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and longer-term samples. All specifications include first-wave fixed effects, and 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 B.12: Treatment effects on Attribution indices (Short term vs Long term)



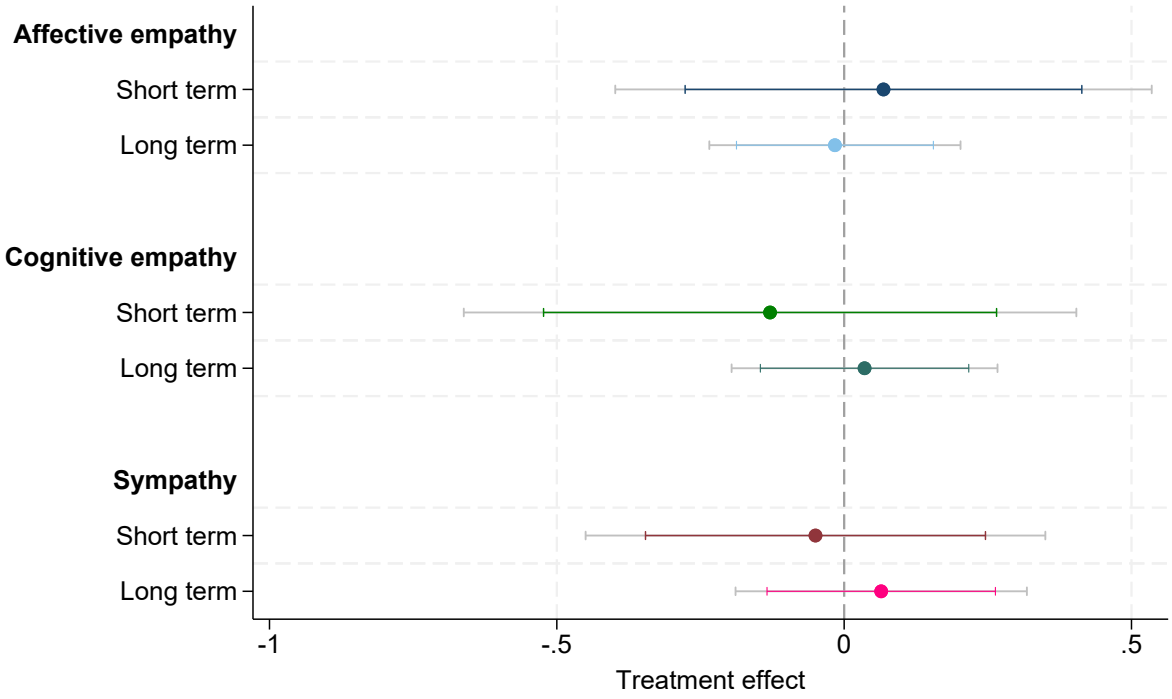
Notes: The figure reports treatment-effect estimates from separate baseline regressions of attribution outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and longer-term samples. Outcomes include the overall attribution index and the seven 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 B.13: Treatment effects on Theory of Mind (Short term vs Long term)



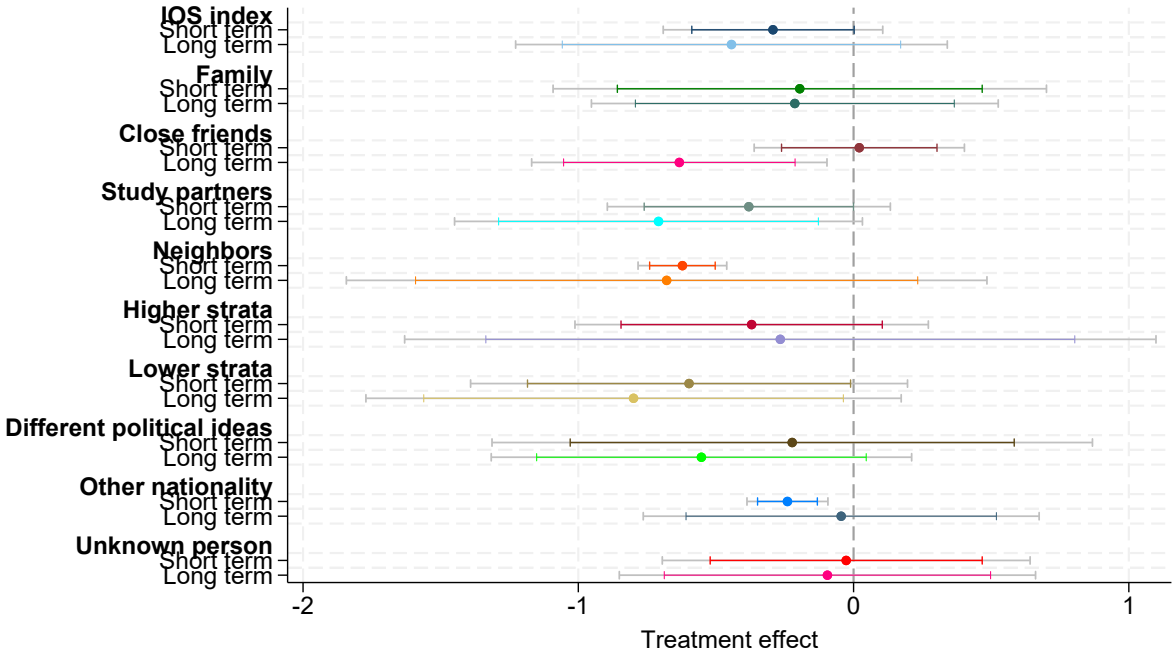
Notes: The figure reports treatment-effect estimates from separate baseline regressions of RMET accuracy on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and longer-term samples. All specifications include first-wave fixed effects, and standard 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 B.14: Treatment effects on Empathy indices (Short term vs Long term)



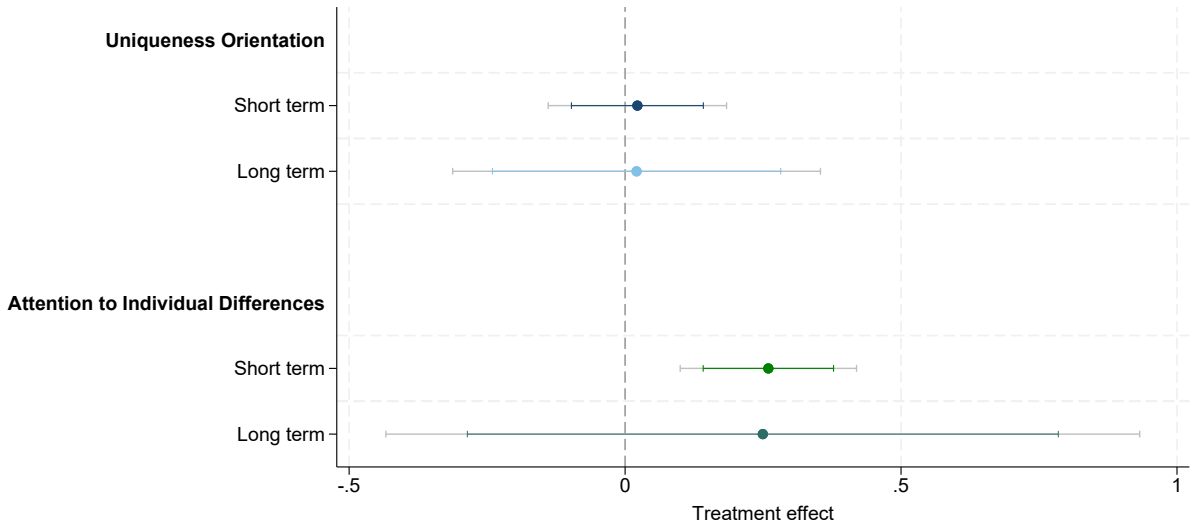
Notes: The figure reports treatment-effect estimates from separate baseline regressions of empathy outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and longer-term samples. All specifications include first-wave fixed effects, and standard 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 B.15: Treatment effects on IOS indices (Short term vs Long term)



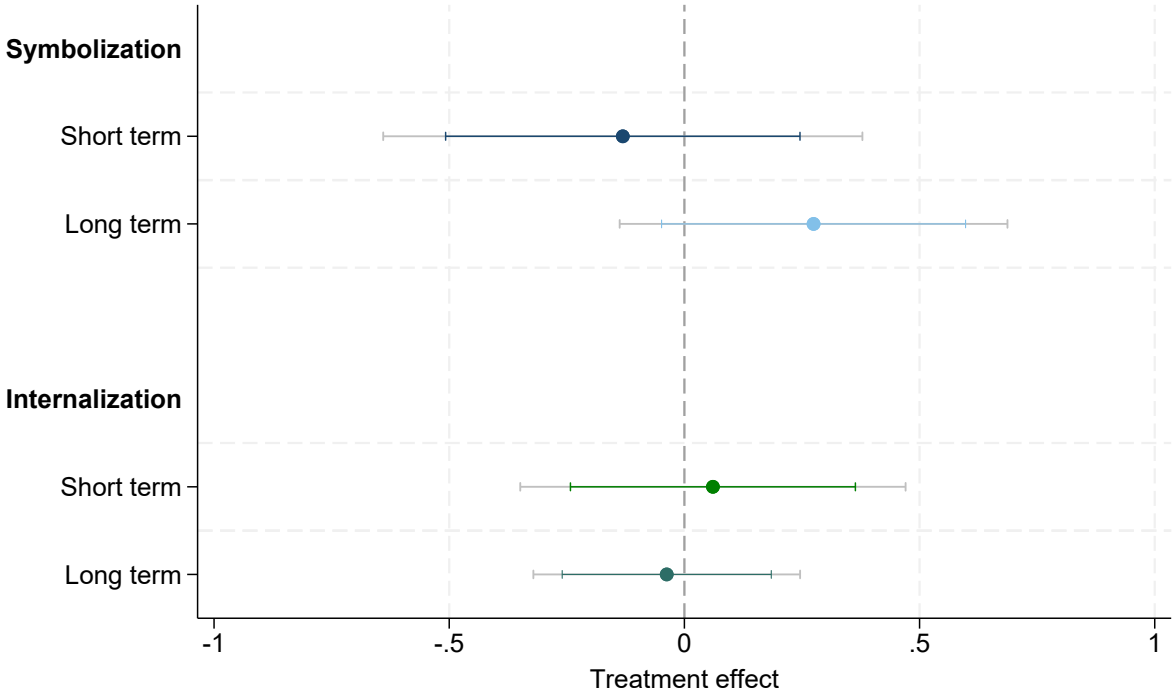
Notes: The figure reports treatment-effect estimates from separate baseline regressions of IOS outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and longer-term samples. Outcomes include the overall IOS index and its component items. All 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 B.16: Treatment effects on Uniqueness (Short term vs Long term)



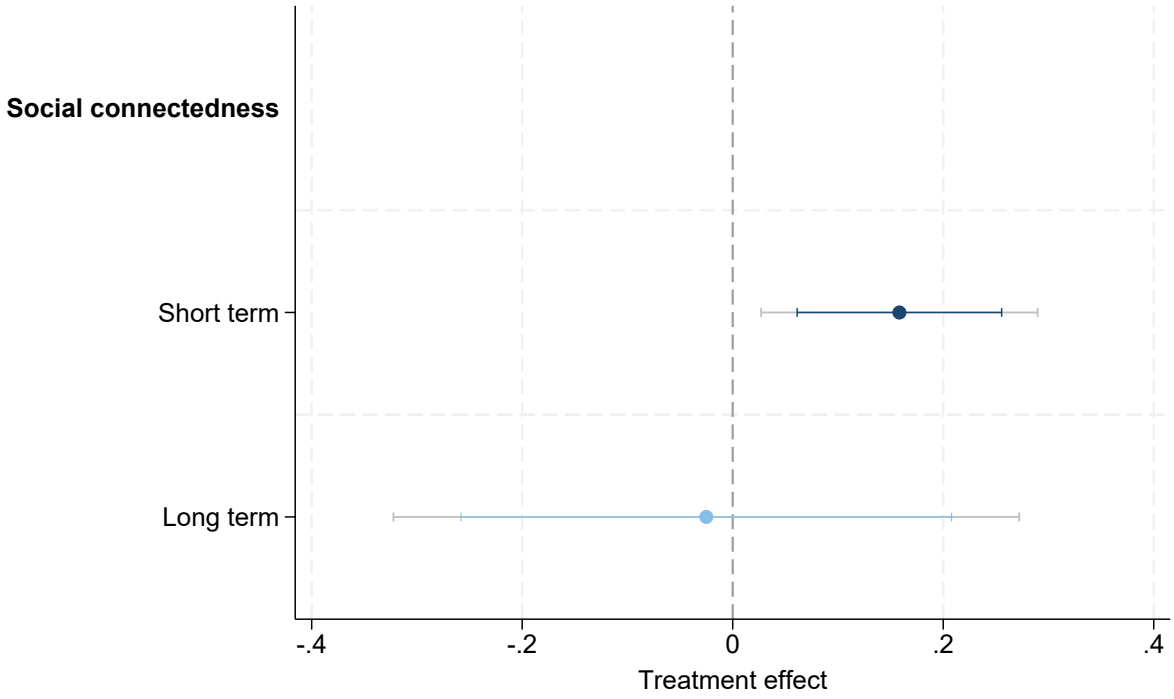
Notes: The figure reports treatment-effect estimates from separate baseline regressions of uniqueness outcomes on an indicator for assignment to the fiction-reading treatment, estimated separately for the short-term and longer-term samples. All specifications include first-wave fixed effects, and standard 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 B.17: 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 longer-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 B.18: 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 longer-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.

B.5.8 Reading-related heterogeneity

This section reports exploratory heterogeneity analyses by baseline reading habits. Because reading behavior is multidimensional, we summarize baseline reading habits using both discrete reading groups and principal component analysis (PCA). The PCA is based on five baseline measures: fiction books read, non-fiction books read, work- or study-related books read, weekly reading hours, and reading frequency. Because the four count-based measures are right-skewed, with a small number of participants reporting unusually high values, we apply a log(1+x) transformation to each before constructing the PCA; reading frequency, which is already measured on a bounded ordinal scale, enters untransformed. This reduces the influence of a handful of extreme readers on the resulting components relative to using raw counts. We retain two components and apply varimax rotation to improve interpretability. The rotated components capture two distinct dimensions of reading behavior. The first component loads mainly on fiction books and weekly reading hours, and is interpreted as fiction-oriented reading. The second component loads mainly on non-fiction and work/study-related books, and is interpreted as instrumental reading. These analyses should be interpreted as exploratory, since baseline reading habits are not experimentally assigned. PCA-based heterogeneity analyses are restricted to participants with complete baseline reading information.

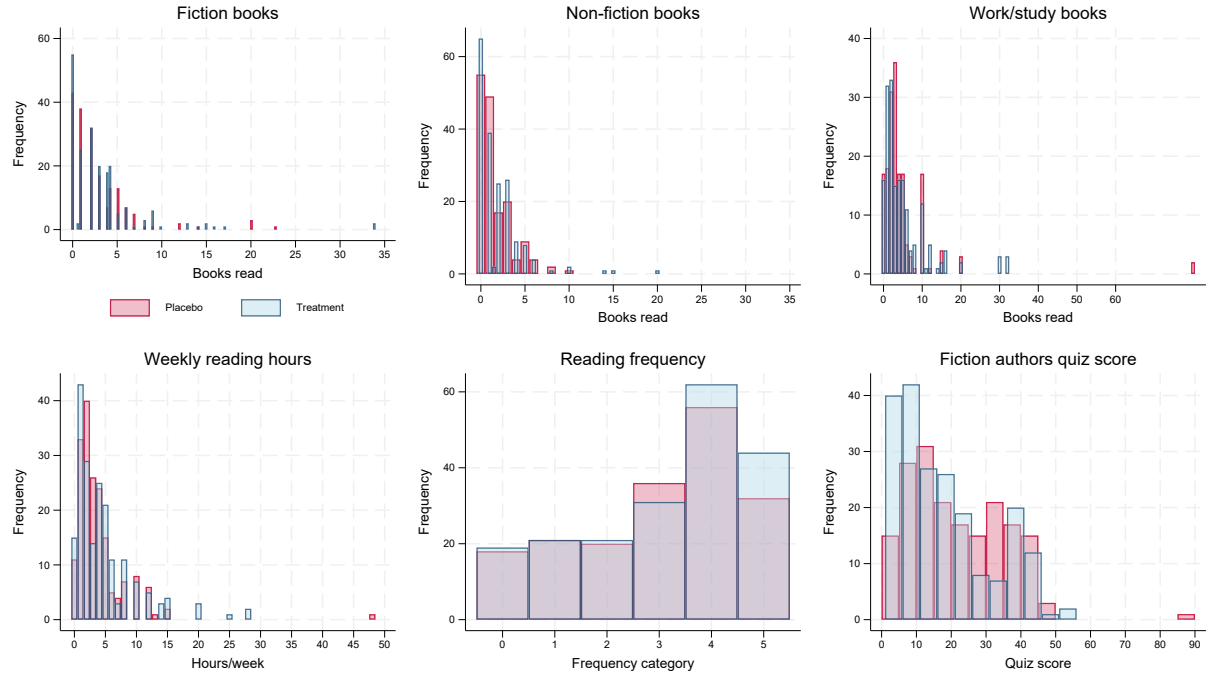


Figure B.19: Distribution of reading-related variables by experimental assignment

Figure B.20 shows the variance explained by each principal component of the log-transformed reading measures. The first two components jointly capture a substantial share of variation

in baseline reading habits, while allowing us to distinguish between fiction-oriented and instrumental reading profiles.

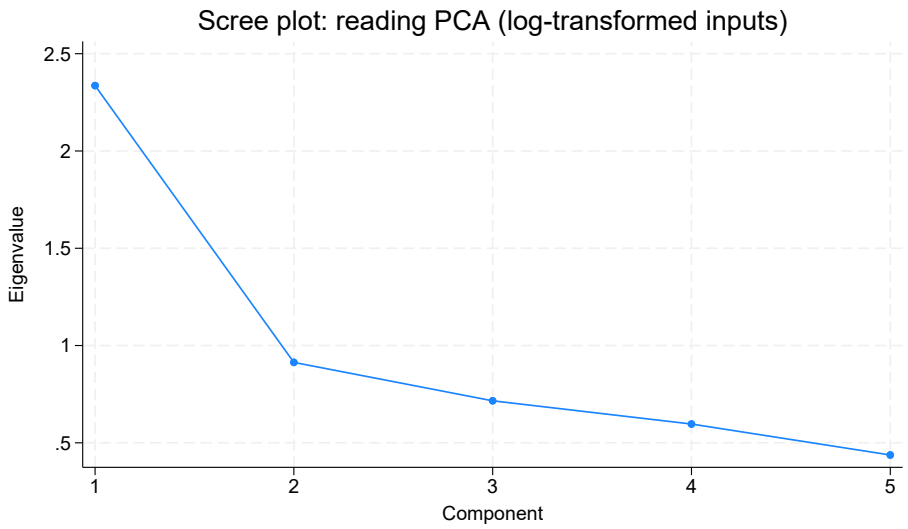


Figure B.20: Scree plot for the PCA of reading-related variables (log-transformed inputs)

Tables B.35 and B.36 report treatment heterogeneity by each rotated PCA component separately. This allows us to examine whether treatment effects vary with fiction-oriented reading and with instrumental reading. Table B.37 reports both rotated components jointly in the same specification, which allows us to assess whether each dimension of reading is associated with treatment heterogeneity net of the other.

Table B.35: Heterogeneity by fiction-oriented reading

	(1) Dictator	(2) Trustworthiness	(3) Ultimatum	(4) Narrative presence	(5) IOS	(6) Attention to individual differences
Treatment	0.509**	0.082***	0.139	0.363**	-0.329**	0.206*
	(0.194)	(0.030)	(0.198)	(0.123)	(0.129)	(0.098)
Fiction-oriented reading	0.002	-0.011	0.059	0.095**	-0.184*	-0.166**
	(0.086)	(0.009)	(0.109)	(0.043)	(0.092)	(0.057)
Treatment × fiction-oriented reading	-0.017	0.028*	-0.181	-0.087	0.130	0.200*
	(0.133)	(0.016)	(0.112)	(0.129)	(0.109)	(0.104)
Constant	3.364***	0.037	4.850***	3.988***	5.678***	3.685***
	(0.846)	(0.061)	(0.556)	(0.390)	(0.532)	(0.416)
Observations	169	1690	169	158	158	158
Control mean	3.683	0.296	4.756	4.888	3.631	3.532
R-squared	0.138	0.426	0.102	0.116	0.152	0.078
First-wave FE	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓

Notes: Standard errors clustered at the session level are reported in parentheses, except for Trustworthiness, which is estimated on long-format trust-game data (one observation per conditional decision, strategy method), interacted with the amount sent by Player 1 and clustered at the participant level; its Observations count and Treatment coefficient are therefore not directly comparable to the other columns. Fiction-oriented reading is obtained from principal component analysis with varimax rotation on log(x+1)-transformed reading counts and loads mainly on fiction books and weekly reading hours. PCA-based heterogeneity analyses are restricted to participants with complete baseline reading information. All specifications include first-wave fixed effects and baseline controls. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.36: Heterogeneity by instrumental reading

	(1) Dictator	(2) Trustworthiness	(3) Ultimatum	(4) Narrative presence	(5) IOS	(6) Attention to individual differences
Treatment	0.506**	0.083***	0.180	0.391***	-0.325**	0.201*
	(0.209)	(0.030)	(0.180)	(0.117)	(0.129)	(0.096)
Instrumental reading	-0.016	-0.003	-0.075	0.269**	-0.122	-0.081
	(0.101)	(0.010)	(0.149)	(0.094)	(0.085)	(0.095)
Treatment × instrumental reading	-0.245*	0.016	0.087	-0.142	-0.067	0.040
	(0.125)	(0.013)	(0.155)	(0.101)	(0.108)	(0.100)
Constant	3.515***	0.025	4.716***	4.064***	6.075***	3.736***
	(0.816)	(0.060)	(0.556)	(0.243)	(0.562)	(0.427)
Observations	169	1690	169	158	158	158
Control mean	3.683	0.296	4.756	4.888	3.631	3.532
R-squared	0.154	0.418	0.093	0.158	0.159	0.057
Observations	155	1250	157	146	148	146
Control mean	3.662	0.289	4.736	4.889	3.598	3.529
R-squared	0.171	0.420	0.107	0.165	0.152	0.069
First-wave FE	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓

Notes: Standard errors clustered at the session level are reported in parentheses, except for Trustworthiness, which is estimated on long-format trust-game data (one observation per conditional decision, strategy method), interacted with the amount sent by Player 1 and clustered at the participant level; its Observations count and Treatment coefficient are therefore not directly comparable to the other columns. Instrumental reading is obtained from principal component analysis with varimax rotation on log(x+1)-transformed reading counts and loads mainly on non-fiction and work/study-related books. PCA-based heterogeneity analyses are restricted to participants with complete baseline reading information. All specifications include first-wave fixed effects and baseline controls. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.37: Heterogeneity by fiction and instrumental reading

	(1) Dictator	(2) Trustworthiness	(3) Ultimatum	(4) Narrative presence	(5) IOS	(6) Attention to individual differences
Treatment	0.495**	0.082***	0.156	0.396***	-0.304**	0.198*
	(0.205)	(0.030)	(0.179)	(0.124)	(0.125)	(0.100)
Fiction-oriented reading	0.019	-0.013	0.090	-0.028	-0.156*	-0.174***
	(0.142)	(0.008)	(0.078)	(0.068)	(0.083)	(0.055)
Instrumental reading	-0.025	0.004	-0.109	0.285**	-0.069	0.021
	(0.160)	(0.009)	(0.139)	(0.119)	(0.070)	(0.109)
Treatment × fiction-oriented reading	0.056	0.028	-0.257**	0.001	0.172	0.224
	(0.177)	(0.017)	(0.107)	(0.148)	(0.097)	(0.124)
Treatment × instrumental reading	-0.268	0.002	0.210	-0.144	-0.123	-0.081
	(0.188)	(0.014)	(0.168)	(0.128)	(0.109)	(0.134)
Constant	3.582***	0.033	4.743***	4.018***	5.946***	3.753***
	(0.713)	(0.061)	(0.637)	(0.320)	(0.557)	(0.396)
Observations	169	1690	169	158	158	158
Control mean	3.683	0.296	4.756	4.888	3.631	3.532
R-squared	0.155	0.427	0.110	0.159	0.174	0.080
First-wave FE	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓

Notes: Standard errors clustered at the session level are reported in parentheses, except for Trustworthiness, which is estimated on long-format trust-game data (one observation per conditional decision, strategy method), interacted with the amount sent by Player 1 and clustered at the participant level; its Observations count and Treatment coefficient are therefore not directly comparable to the other columns. Fiction-oriented reading and instrumental reading are obtained jointly from principal component analysis with varimax rotation on log(x+1)-transformed reading counts and are included together in the same specification; fiction-oriented reading loads mainly on fiction books and weekly reading hours, while instrumental reading loads mainly on non-fiction and work/study-related books. PCA-based heterogeneity analyses are restricted to participants with complete baseline reading information. All specifications include first-wave fixed effects and baseline controls. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Figures B.21 and B.22 visualize heterogeneity by instrumental reading. This dimension is particularly relevant because it distinguishes participants whose baseline reading habits were more academic or work-related from those whose reading profile was less instrumental.

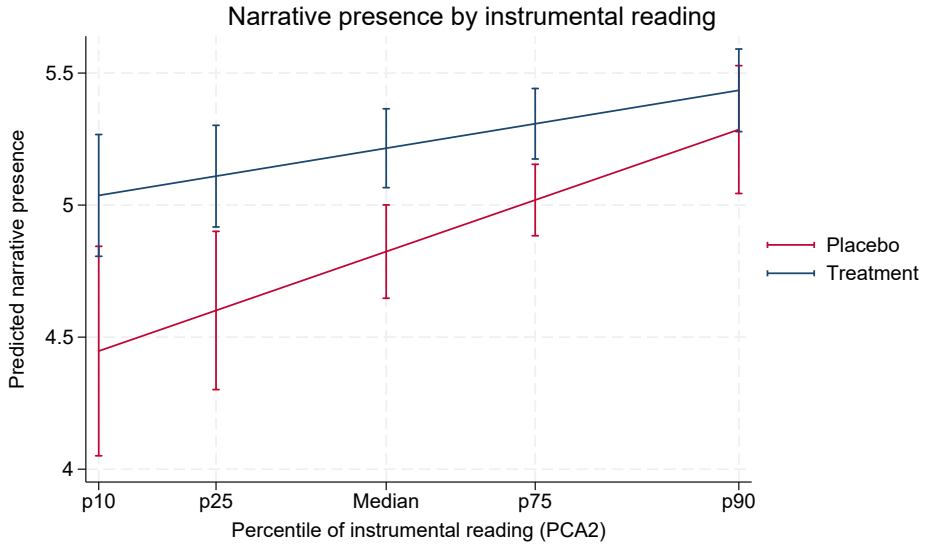


Figure B.21: Narrative presence by instrumental reading

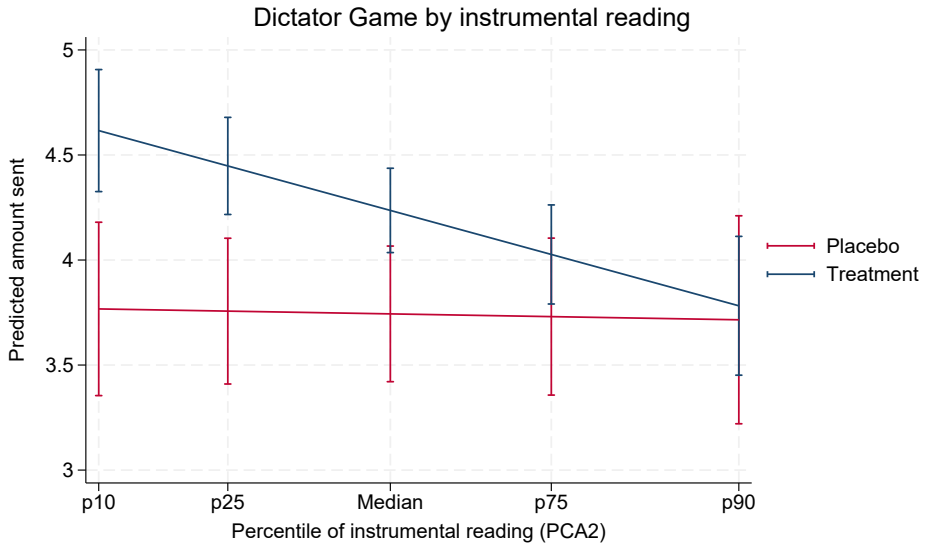


Figure B.22: Dictator Game transfers by instrumental reading

Figures B.23 and B.24 provide the analogous visualization for fiction-oriented reading. These figures help assess whether the effects of the fiction course are stronger among participants who were already more oriented toward fiction and narrative reading at baseline.

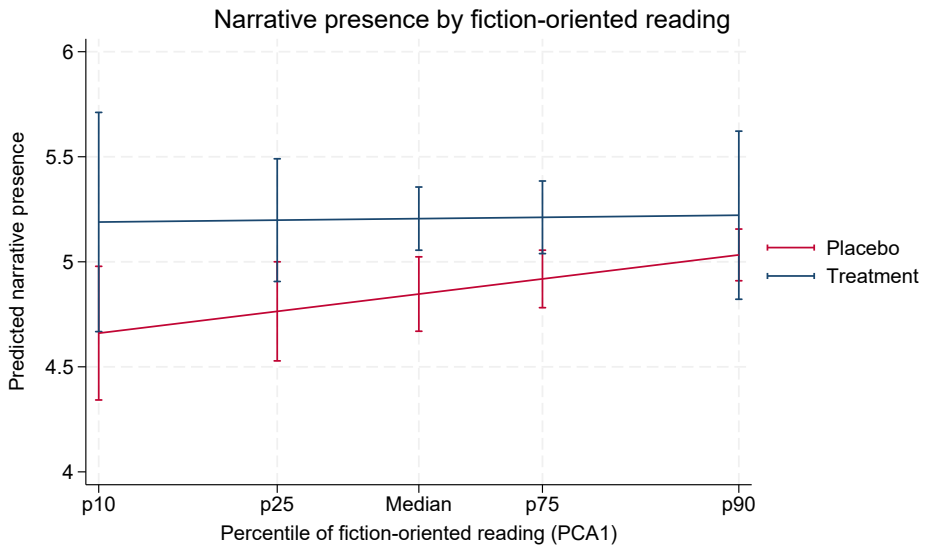


Figure B.23: Narrative presence by fiction-oriented reading

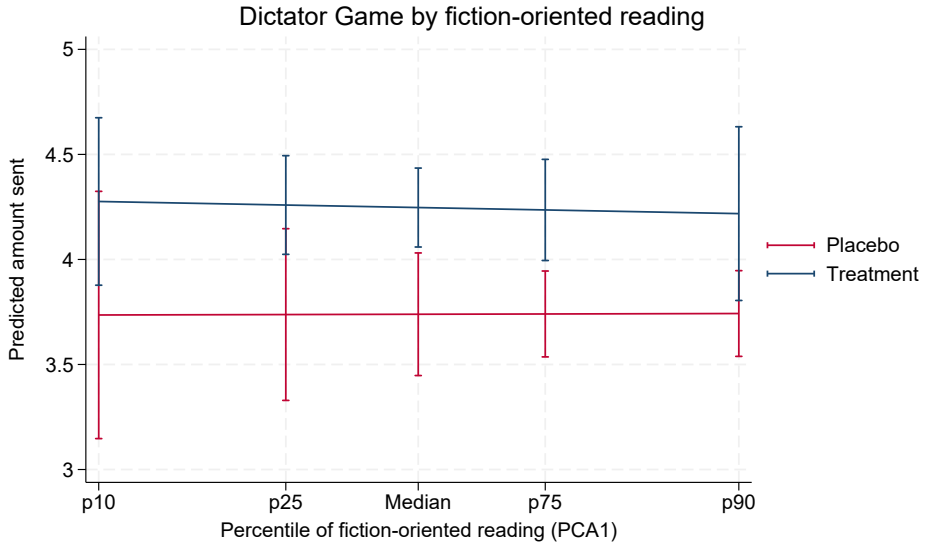


Figure B.24: Dictator Game transfers by fiction-oriented reading

Figures B.25 and B.26 summarize both reading dimensions in a single plot. For each outcome, we overlay the estimated treatment effect across percentiles of fiction-oriented reading (PCA1) and of instrumental reading (PCA2), holding the other component fixed at its sample mean, from the same joint specification as Table B.37. Because the two components are on different scales, the horizontal axis reports the percentile rank of each component (p10, p25, median, p75, p90) rather than its raw value, so the two curves can be compared directly on a common axis.

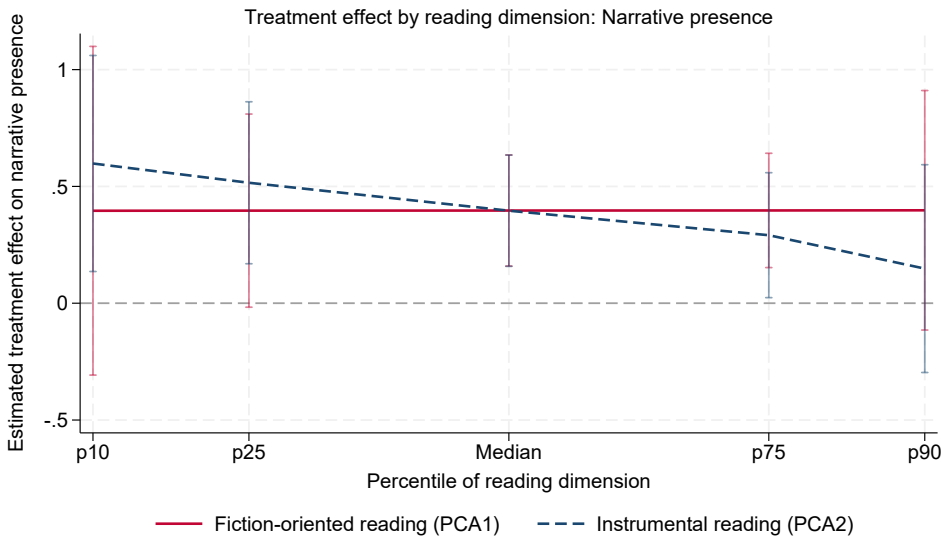


Figure B.25: Narrative presence: treatment effect by fiction-oriented vs. instrumental reading

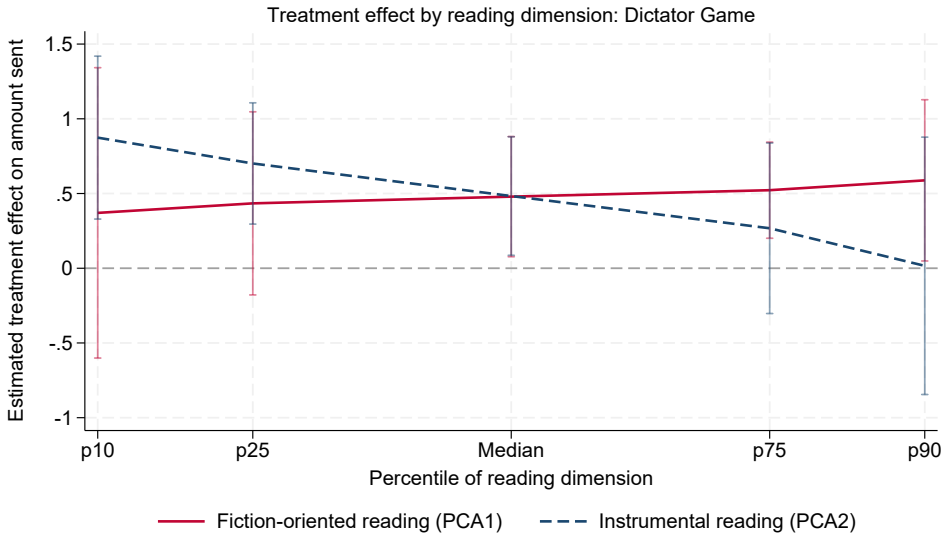


Figure B.26: Dictator Game: treatment effect by fiction-oriented vs. instrumental reading

B.5.9 Additional psychological results

Table B.38: Narrative engagement: complete set of 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	0.307**	0.377**	0.080	0.132	0.234*	0.313***	0.341***	0.397***	0.164*	0.228**	0.218	0.334**	0.425***	0.508***	0.176	0.279
	(0.132)	(0.139)	(0.126)	(0.106)	(0.115)	(0.090)	(0.089)	(0.103)	(0.083)	(0.084)	(0.122)	(0.144)	(0.104)	(0.105)	(0.155)	(0.179)
Constant	4.060***	3.456***	4.855***	5.774***	4.943***	4.437***	4.830***	3.814***	5.198***	3.456***	4.248***	2.157**	4.482***	3.428***	4.940***	1.186*
	(0.050)	(0.477)	(0.073)	(0.803)	(0.060)	(0.682)	(0.063)	(0.415)	(0.068)	(0.489)	(0.088)	(0.764)	(0.084)	(1.013)	(0.131)	(0.603)
Placebo mean	4.055	4.055	4.853	4.853	4.945	4.945	4.824	4.824	5.188	5.188	4.243	4.243	4.478	4.478	4.945	4.945
Observations	175	175	175	175	175	175	175	175	175	175	175	175	175	175	175	175
R^2	0.063	0.123	0.004	0.036	0.016	0.052	0.078	0.132	0.117	0.178	0.026	0.120	0.038	0.073	0.020	0.130
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 B.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 B.39: Attributional complexity: complete set of subscales

	Attribution (mean)		Motivation		Complex		Metacognitive		Behavior		Ambiguity		Situational		Cognitive effort	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Treatment	0.153	0.155	0.122	0.117	0.134	0.177**	0.294	0.277	0.048	0.013	0.179*	0.188*	0.167	0.165	0.127	0.149
	(0.087)	(0.095)	(0.130)	(0.129)	(0.076)	(0.059)	(0.172)	(0.175)	(0.093)	(0.089)	(0.084)	(0.089)	(0.112)	(0.128)	(0.097)	(0.097)
Constant	5.018***	4.906***	4.854***	4.943***	5.208***	5.662***	4.984***	4.583***	5.387***	4.648***	4.884***	5.189***	4.804***	4.816***	5.004***	4.505***
	(0.035)	(0.504)	(0.060)	(0.756)	(0.055)	(0.384)	(0.069)	(0.964)	(0.034)	(0.750)	(0.044)	(0.614)	(0.048)	(0.592)	(0.071)	(0.566)
Placebo mean	5.022	5.022	4.859	4.859	5.212	5.212	4.991	4.991	5.394	5.394	4.885	4.885	4.806	4.806	5.009	5.009
Observations	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184
R^2	0.029	0.098	0.016	0.034	0.015	0.093	0.036	0.080	0.027	0.094	0.013	0.054	0.010	0.096	0.015	0.112
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 the 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 greater attributional complexity. Columns (3)–(16) report results for the seven attributional complexity subscales: *Motivation* (items 1, 8, 15, and 22), *Complex* (items 2, 9, 16, and 23), *Metacognitive* (items 3, 10, 17, and 24), *Behavior* (items 4, 11, 18, and 25), *Ambiguity* (items 5, 12, 19, and 26), *Situational* (items 6, 13, 20, and 27), and *Cognitive effort* (items 7, 14, 21, and 28). Each index is computed as a respondent-level mean over non-missing items. 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 B.7 for the full list of items and coding details.

Table B.40: Inclusion of the Other in the Self: complete results

	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	-0.379**	-0.357**	-0.176	-0.154	-0.262*	-0.284*	-0.565***	-0.531***	-0.675***	-0.650***	-0.421*	-0.315	-0.658***	-0.663***	-0.366	-0.389*	-0.205	-0.166	-0.082	-0.066
	(0.127)	(0.133)	(0.189)	(0.192)	(0.141)	(0.144)	(0.142)	(0.150)	(0.184)	(0.177)	(0.222)	(0.232)	(0.180)	(0.183)	(0.205)	(0.212)	(0.139)	(0.124)	(0.155)	(0.156)
Constant	3.619***	5.401***	5.345***	8.700***	5.174***	7.231***	4.141***	5.768***	2.994***	5.041***	3.487***	5.065***	3.740***	5.996***	2.588***	4.447***	3.284***	3.887***	1.816***	2.471**
	(0.103)	(0.640)	(0.132)	(0.992)	(0.109)	(0.852)	(0.133)	(0.659)	(0.162)	(1.029)	(0.199)	(1.218)	(0.108)	(0.952)	(0.168)	(1.111)	(0.130)	(0.894)	(0.120)	(0.797)
Placebo mean	3.616	3.616	5.341	5.341	5.165	5.165	4.141	4.141	2.988	2.988	3.482	3.482	3.741	3.741	2.588	2.588	3.282	3.282	1.812	1.812
Observations	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184	184
R^2	0.046	0.100	0.006	0.071	0.021	0.060	0.039	0.067	0.062	0.088	0.022	0.094	0.050	0.096	0.021	0.090	0.006	0.067	0.009	0.028
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 B.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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.41: Empathy indices: complete results

	Affective		Cognitive		Sympathy	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.024	0.056	-0.032	-0.022	0.023	0.028
	(0.076)	(0.086)	(0.093)	(0.099)	(0.080)	(0.085)
Constant	5.207***	4.079***	5.313***	5.818***	4.722***	3.905***
	(0.066)	(0.751)	(0.042)	(0.797)	(0.043)	(0.559)
Control mean	5.198	5.198	5.306	5.306	4.710	4.710
Observations	175	175	175	175	175	175
R^2	0.119	0.173	0.064	0.085	0.246	0.281
Wave fixed effects	✓	✓	✓	✓	✓	✓
Controls		✓		✓		✓

Notes: Entries report coefficients from linear regressions of empathy outcomes on an indicator for assignment to the treatment condition. The dependent variables are respondent-level indices constructed from post-reading narrative-response items. *Affective empathy* is the mean of items 1–5, with item 4 reverse-coded, and captures emotional resonance with others' feelings. *Sympathy* is the mean of items 6–8 and captures concern or compassion for others. *Cognitive empathy* is the mean of items 9–14, with items 12 and 13 reverse-coded, and captures perspective-taking and understanding of others' emotions and situations. Higher values indicate greater empathy in each domain. All specifications include first-wave fixed effects. Even-numbered columns additionally include demographic controls. See Appendix Table B.7 for the full list of items and coding details. Standard errors clustered at the session level are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.42: RMET index

	(1)	(2)
Treatment	0.158	0.192
	(0.226)	(0.250)
Constant	6.396***	5.339***
	(0.190)	(0.974)
Placebo mean	6.424	6.424
Observations	183	183
R^2	0.141	0.186
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 B.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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.43: Social connectedness

	(1)	(2)
Treatment	0.090	0.072
	(0.060)	(0.061)
Constant	2.305***	2.880***
	(0.046)	(0.311)
Placebo mean	2.304	2.304
Observations	175	175
R^2	0.009	0.074
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 B.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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B.44: Morality indices

	Symbolization		Internalization	
	(1)	(2)	(3)	(4)
Treatment	0.003	0.064	-0.002	0.002
	(0.155)	(0.120)	(0.065)	(0.078)
Constant	4.237***	3.020***	5.396***	4.926***
	(0.136)	(0.707)	(0.044)	(0.655)
Placebo mean	4.233	4.233	5.395	5.395
Observations	175	175	175	175
R^2	0.012	0.104	0.000	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 B.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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.