Maastricht University

ROA

Group identity and betrayal: decomposing trust

Maria Polipciuc

ROA Research Memorandum

ROA-RM-2022/2

Researchcentrum voor Onderwijs en Arbeidsmarkt | ROA
Research Centre for Education and the Labour Market | ROA

Group identity and betrayal: decomposing trust

Maria Polipciuc

ROA-RM-2022/2
May 2022

Research Centre for Education and the Labour Market
Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands
T +31 43 3883647

secretary-roa-sbe@maastrichtuniversity.nl
www.roa.nl

ISSN: 2666-8823

Abstract

Group identity and betrayal: decomposing trust[*],[**]

Betrayal aversion is an important factor in the decision to trust. Trust in members of one's own social group (ingroup members) is often higher than that in members of other groups (outgroup members). In this paper, I study (i) how betrayal aversion contributes to in-/outgroup discrimination in trust and (ii) how this contribution evolves as social groups solidify.

I run two very similar laboratory experiments, first shortly after individuals have been randomly assigned to social groups (outside the laboratory), and seven months later. I find a null result: there is no intergroup discrimination in betrayal aversion, at neither point in time. In the first experiment, betrayal aversion is positive, and does not differ towards in- versus outgroup members. In the second experiment, I find no betrayal aversion. At this time, a subsample of participants trusts ingroup members more, but only in the first of two trusting decisions they make. Factors other than betrayal aversion—such as beliefs about trustworthiness and outcome-based social preferences—seem to explain this ingroup bias in trust.

I suggest a couple of potential explanations for the lack of betrayal aversion in the second experiment.

JEL classification: C72, C91, J15, J16
Keywords: Trust, betrayal aversion, social identity, experiment

Maria Polipciuc
Vienna University of Business and
Economics and Maastricht University
Institute for Markets and Strategy
Welthandelsplatz 1, Building D5
1020 Vienna
maria.polipciuc@wu.ac.at

[*] Go to https://doi.org/10.17026/dans-xyw-9e4r to visit the page for additional materials.

[**] First and foremost, I thank Martin Strobel for his generous and kind supervision. Raymond Montizaan, Frauke Meyer, and Frauke Stehr offered valuable feedback. I thank seminar audiences at Maastricht University, RWTH Aachen University, conference participants at the 2018 ESA World Meeting in Berlin and at the 2018 M-BEES in Maastricht. I also thank Kelly Geyskens for the opportunity to run the experiments. Joey Mak was an extremely helpful contact at the Administration Office. Miruna Coteţ, Alexandra Franzen, Ferdinand Pieroth and Lars Wittrock provided excellent research assistance with running the experiments. Discussions with Nickolas Gagnon and Henrik Zaunbrecher helped shape the ideas in this paper. The author gratefully acknowledges support from an A.SK Fellowship.

1 Introduction

Trust is essential for economic and social interactions. Many contracts are incomplete and difficult to enforce, and thus depend crucially on trust (Arrow, 1974; Schwerter and Zimmermann, 2020). Trust is positively correlated with a host of economic outcomes, ranging from the macro level, such as economic growth (La Porta et al., 1997; Zak and Knack, 2001) or volume of international trade (Guiso et al., 2009), to the personal level, such as personal income (Butler et al., 2016).

There is ample evidence that social distance influences trust. Most experimental studies find homophily (ingroup bias and/or outgroup discrimination) in trust: individuals trust members of their own group (ingroup members) more than members of other groups (outgroup members) (Glaeser et al., 2000; Hargreaves Heap and Zizzo, 2009; Etang et al., 2010; Brandts and Charness, 2011; Guillen and Ji, 2011; Binzel and Fehr, 2013; Chuah et al., 2013; Falk and Zehnder, 2013).[1] One way to examine why this is the case is to look at how social distance influences the determinants of trust. These can be split into individual determinants (beliefs and preferences) and institutional determinants (features of the environment in which the interaction takes place).

In this paper, I focus on an individual determinant, betrayal aversion, which many studies find to be important for trusting decisions (Bohnet and Zeckhauser, 2004; Aimone et al., 2015; Fairley et al., 2016; Quercia, 2016; Bacine and Eckel, 2018; Butler and Miller, 2018; Polipciuc and Strobel, 2022, this list is not exhaustive). Aimone et al. (2015) define betrayal aversion as "disutility from the experience, anticipation or observation of non-reciprocated trust". Later studies show betrayal aversion is a preemptive reaction to the perceived (malevolent) intentions of the opponent (Butler and Miller, 2018; Polipciuc and Strobel, 2022). Based on these results, in this paper I consider betrayal aversion to be an intention-based social preference. Intention-based social preferences are one of the types of individual determinants of trust (Cox, 2004; Fehr, 2009; Stanca et al., 2009; Strassmair, 2009; Johnsen and Kvaløy, 2016), together with beliefs about the opponent's trustworthiness (Ashraf et al., 2006; Sapienza et al., 2013; Costa-Gomes et al., 2014) and

[1] There is however large variation depending on the type of group identity used. Some studies do not find in-/outgroup discrimination in trust (Fershtman and Gneezy, 2001; Güth et al., 2008). For more details, see a survey of the economics literature on group identity and discrimination in trust and trustworthiness of the last 20 years (Li, 2020, p. 11–12) and the meta-analyses of discrimination in experiments by Balliet et al. (2014) and Lane (2016).

Lane (2016) finds that in-/outgroup discrimination in economic games is strongest when individuals belong to socially or geographically distinct groups. Balliet et al. (2014)—who include experiments across the social sciences, but whose inclusion criteria drop two-thirds of the economics studies in Lane (2016)— find that discrimination by trust game senders is larger than that by dictators in dictator games.

outcome-based social preferences (Cox, 2004; Ashraf et al., 2006; Sapienza et al., 2013).[2] While there is evidence of ingroup bias in beliefs about trustworthiness and outcome-based social preferences (see Li, 2020, for a review), I am aware of only one study examining the influence of social distance on betrayal aversion: Bacine and Eckel (2018), which I present in subsection 2.3.

I use two laboratory experiments to study how trust and betrayal aversion vary with the identity of the opponent (ingroup/outgroup) in a binary trust game (Bohnet and Zeckhauser, 2004). The first experiment takes place a month after groups were formed through random assignment (at T1), and the second one seven months later (at T2).[3] This way, I can measure the impact of betrayal aversion to in-/outgroup members on trust when identity is new and carries little meaning and later on, when it is more defined. I find positive, non-discriminatory trust and betrayal aversion at T1. At T2, there is no discrimination in trust in the aggregate, and betrayal aversion is not significantly different from zero (and non-discriminatory). Exploratory analysis at T2 shows that a subgroup of trustors has an ingroup bias in trust in the first of the two decisions they make. This stems not from differential betrayal aversion, but from differences in a component which jointly measures risk aversion, reactions to beliefs about trustworthiness, and outcome-based social preferences. As a result, in neither experiment is there discrimination in betrayal aversion.

This paper has three main contributions. First, it adds to the literature on individual determinants of intergroup discrimination in trust. The results in this paper suggest that risk preferences, beliefs about trustworthiness and outcome-based social preferences (such as altruism) play a bigger role than betrayal aversion in the decision to trust.

Second, in light of the null result at T2, where I did not find betrayal aversion, this paper raises the question whether betrayal aversion is robust to a more stringent identification like the one used in this paper (the design was adapted from Aimone and Houser, 2012). I identify betrayal aversion as a residual, after ensuring individual trustors' subjective beliefs are constant across treatments within an opponent type (in- or outgroup). This avoids potential confounding factors should trustors not be expected utility maximizers.[4] The original design used by most papers which find betrayal aversion is incentive-compatible under the assumption that trustors do not violate the Substitution Axiom of expected utility. However, this violation has been shown empirically to be rather common (Starmer, 2000;

[2] Some studies find risk attitudes to be a determinant of trust e.g. in the treatment comparison approach in citeEngelmann2021, while others do not e.g. Ashraf et al. (2006).

[3] The groups—which are social groups of students—exist outside the lab. Random assignment to a group is done independently of this study. More details about the setting and the groups are available in Appendix A.

[4] I present the design in detail in Section 3.

Li et al., 2020, p. 275). Future studies should replicate the findings on betrayal aversion using this more stringent definition.

Third, the strategy used in this paper to measure betrayal aversion is potentially useful for other experimental studies wishing to disentangle statistical from taste-based discrimination in settings with uncertainty.[5] Bohren et al. (2019) show that one can properly identify these components when there is uncertainty by designing a control treatment which keeps *subjective* beliefs constant (instead of *objective* probabilities of the opponent's behavior). Here, since betrayal aversion towards an opponent type is identified while keeping subjective beliefs constant across treatments, it corresponds to the intention-based portion of taste-based discrimination.[6]

I conclude that there is no evidence of taste-based discrimination in trust due to intention-based social preferences, neither at T1, nor at T2. With time, a subgroup of individuals trust ingroup members more. This is driven by an increase over time in a component reflecting statistical discrimination combined with taste-based discrimination due to outcome-based social preferences.

The paper is structured as follows. Section 2 presents the related literature. Section 3 describes the experimental design. Section 4 explains the conceptual framework and presents the hypotheses. Section 5 presents the data and the results and Section 6 concludes.

2 Related literature

This paper is mostly related to two strands of literature: the literature on betrayal aversion and the literature on in-/outgroup discrimination in intention-based social preferences.[7] In this section, I first describe how betrayal aversion has been

[5] Economists usually distinguish between statistical discrimination and taste-based discrimination. Statistical discrimination is the part of discrimination which is rational and it is based on beliefs about the opponent's behavior given her group identity (Arrow, 1973). Taste-based discrimination is the part of discrimination which is not responsive to a change in beliefs. It is attributed to preferences (Becker, 2010).

This distinction is useful from a practical point of view. For instance, providing information about the frequency of a certain behavior is potentially successful in addressing statistical discrimination if the actual behavior of members of a group is cooperative more frequently than expected.

[6] This is similar to identifying taste-based discrimination as a residual after controlling for beliefs about in- and outgroup members, a widely used method (see Lane, 2016, footnote 22). The innovative aspect is that the adapted design controls for *subjective* beliefs, such that potential differences between subjective beliefs and objective probabilities do not "contaminate" the measure of taste-based discrimination. For a detailed explanation, see Section 3.

[7] Intentions are a type of conditional social preferences. Conditional social preferences—which are potentially relevant in strategic interactions—have been modeled either using psychologi-

identified in Bohnet and Zeckhauser (2004) and potential issues with this design. Next, I present the literature on betrayal aversion. In the last part of this section, I summarize findings about discriminatory behavior of trust game senders in laboratory experiments.

2.1 Identifying betrayal aversion

The term "betrayal aversion" was introduced by Bohnet and Zeckhauser (2004) (henceforth BZ). BZ use a modified version of the trust game (Berg et al., 1995) called *the binary trust game*. The game used by BZ is presented in Figure 1. In this version, a first mover (he) has to choose between a safe option (*Out*) and a risky option (*In*). The risky option increases efficiency (the total payoff available in the game), but may lead to a higher or to a lower payoff for the first mover than the safe option. This depends on whether the second mover shares the multiplied amount equally (*Left*)—which means she returns an amount greater than the amount the first mover had sent her—or keeps most of it for herself (*Right*). The first mover's option *In* is interpreted as him trusting the second mover. The second mover's option *Left* is interpreted as her returning the first mover's trust.

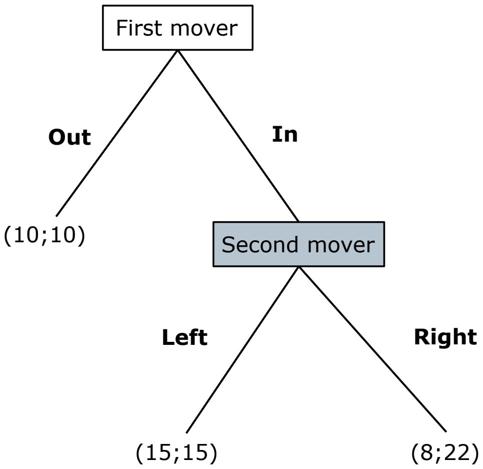

Figure 1: The binary trust game in Bohnet and Zeckhauser (2004)

To identify betrayal aversion, BZ compare first mover behavior in the binary trust game with a dictator's behavior in a control dictator game, dubbed "the risky dictator game". The risky dictator game is identical to the binary trust game,

cal game theory, by incorporating higher order beliefs into the utility function (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004) or using the "revealed intentions" approach (Cox et al., 2007, 2008). Falk and Fischbacher (2006) and the general model in Charness and Rabin (2002) combine outcome-based preferences and intentions-based social preferences.

except for one thing: at the second node, the decision is made by a randomization device. Unbeknownst to players, the probability that the equal split is implemented in the risky dictator game by the randomization device, p^*, is the same as the probability that a randomly chosen second mover chooses the equal split in the binary trust game. Thus, the two games are equally risky, but differ in the source of risk: a human decision in the binary trust game, and a random event in the risky dictator game.

BZ ask first movers in the binary trust game and dictators in the risky dictator game (for simplicity, I will call both types of players "first movers") what their minimum required threshold is for p^* such that for values equal to or above their threshold they prefer the risky option *In* over the safe option *Out*. This value is called a first mover's *minimum acceptable probability*—in short, MAP. If p^* is equal to or above a first mover's MAP, his randomly matched opponent's decision determines the payoffs in the trust game and a random draw from the distribution (*Left*, p^*;*Right*,$1-p^*$) determines it in the risky dictator game. If p^* is below a first mover's MAP, *Out* is implemented. BZ find that participants have lower MAPs on average for taking the risky option when risk is random than when it is strategic. They define betrayal aversion as the positive premium between the average MAP in the binary trust game and the average MAP in the risky dictator game.

This elicitation procedure is similar to the Becker–DeGroot–Marschak (BDM) procedure (Becker et al., 1964). Unlike the standard version of the BDM, in BZ p^*, the value with which a participant's MAP is compared to determine payoffs, is not drawn from a uniform distribution. This means first movers are likely to have a different distribution in mind for p^* in the trust game and in the risky dictator game. In the risky dictator game, since participants are not told how p^* is generated, they most likely assume it to be uniformly distributed (Li et al., 2020). Bohnet et al. (2008, p. 298) and Bohnet et al. (2010, p. 815–816) acknowledge this and argue MAPs elicited this way should not be affected by ambiguity aversion if first movers adhere to the Substitution Axiom of von Neumann-Morgenstern utility. It is thus a normative requirement in BZ that participants be rational expected utility maximizers in order for betrayal aversion to be identifiable using this procedure (Li et al., 2020). Later papers on betrayal aversion can be divided into those which use the original BZ design and thus assume implicitly or explicitly that first movers satisfy this requirement (the large majority of papers) and those which do not.

2.2 Literature on betrayal aversion

Many papers following BZ find evidence of betrayal aversion. Bohnet et al. (2008) replicate the results of BZ in Brazil, China, Oman, Switzerland, Turkey, and the United States. Bohnet et al. (2010) run experiments in Kuwait, Oman, Switzerland, the United States, and the United Arab Emirates, and conclude that cross-regional differences in trust are due to differences in intolerance to betrayal aversion. Aimone and Houser (2011) find that betrayal aversion can be beneficial for trust relationships: trustees who know they are facing a betrayal averse trustor are more likely to return his trust. Aimone and Houser (2012) modify how uncertainty in the game is resolved and show that betrayal aversion is a distinct concept from loss aversion (as the findings in Bohnet et al., 2010, might suggest there is overlap). Aimone et al. (2015) measure betrayal aversion at the individual level. They conclude that individual risk aversion and individual betrayal aversion are uncorrelated. Members of high status groups and those who have an unusual trajectory compared to their peers seem to be more betrayal averse (Hong and Bohnet, 2007; Suchon and Villeval, 2019). Quercia (2016) shows that eliciting betrayal aversion using a multiple price list is easier for subjects to understand, and yields qualitatively similar results with the original elicitation method.

Recent studies support BZ's preferred interpretation that betrayal aversion is a reaction of first movers to the perceived intentions of second movers. Butler and Miller (2018) find that betrayal aversion vanishes or becomes negative when first movers know that their opponents are oblivious to the consequence of their own actions (and thus cannot form intentions). Polipciuc and Strobel (2022) replicate the findings of BZ in a trust game, but find no strategic risk premium (nor a discount) in a game in which the two players' interests are aligned and thus the second mover's motivation is hard to deduce from her actions.

There are also a couple of studies which do not find betrayal aversion or find it to be limited in scope. Fetchenhauer and Dunning (2012) find that respondents prefer to take a risk by trusting someone to placing a bet when the chance of a high payoff from either choice is low (46%), but that they are equally likely to choose either of the two options when the chance of a high payoff is high (80%). In their setup, there is no uncertainty about p^* in either treatment. In a within-subject design, Breuer and Hüwe (2014) find that participants send equal amounts of money in a trust game and when betting in an equiprobable bet. Fetchenhauer and Dunning (2012) and Breuer and Hüwe (2014) ensure participants in the trust game and in the control game had the same distribution of p^* in mind, but neither paper controls for outcome-based social preferences, as they do not include a second player in the control game. Li et al. (2020, Appendix A) show theoretically that a strategic premium in the trust game may occur due to many things other than betrayal aversion, such as "ambiguity attitudes, complexity, *different beliefs*, and dynamic optimization", if first movers are not expected utility maximizers (emphasis added). In a study which uses several control games in order to quantify the importance of risk aversion, beliefs, outcome-based social preferences and betrayal aversion for trusting, Engelmann et al. (2021, personal communication) find that betrayal aversion only seems to contribute as an isolated component when beliefs about trustworthiness are very high (outside the range which is found empirically). Their control games are designed to ensure that p^* is the same across treatments and that this is known to participants.

In conclusion, most studies which keep p^* equal across treatments—and inform participants about it—do not find that betrayal aversion plays a significant role in the decision to trust (with the exception of Aimone and Houser, 2012; Polipciuc and Strobel, 2022).[8] With this in mind, in this study I adapt the design of Aimone and Houser (2012) to cleanly identify betrayal aversion even if first movers are not expected utility maximizers.

2.3 Experimental literature on discrimination in trust

Lane (2016) carries out a meta-analysis of discrimination in laboratory experiments. The study does not present a breakdown by role for studies focusing on identifying statistical versus taste-based discrimination (e.g. how many refer to trust game senders). However, it mentions that from the 60 cases where there is scope for both statistical and taste-based discrimination in papers which aim to disentangle the two, 26 do not find any of the two.[9] From the remainder, results for trust game senders are mixed: there are two cases of both statistical and taste-based discrimination, seven cases of taste-based discrimination only, one case of statistical discrimination only, nine cases of taste-based outgroup favoritism, and one case of statistical outgroup favoritism only (for details, see Table A.3 in Lane, 2016). More often than not, discrimination (or favoritism) by trust game senders seems to have a significant taste-based component.

The recent literature on discrimination in conditional social preferences is summarized in Li (2020, p. 8–9). None of the studies mentioned focuses on first mover behavior in trust games. Bacine and Eckel (2018) is the paper most closely related to mine, also studying betrayal aversion towards in- and outgroup members. The authors find outgroup discrimination in trust and in betrayal aversion. The study design does not ensure constant beliefs across games within an opponent type, so it is not clear which type of discrimination is captured by the premium required to trust outgroup members—nor whether this reflects discrimination in betrayal aversion or in one of the confounds pointed out by Li et al. (2020).

[8] The data collected at T1 for this study is also part of a larger data set used in Polipciuc and Strobel (2022).

[9] Lane (2016) defines a case as one group discriminating against another group. Most studies thus include several cases each.

3 Experimental design

I use variants of the two-player, two-stage binary trust game and risky dictator game from BZ. Payoffs differ, but the equilibrium structure is the same. Figure 2 presents the two treatments. Payoffs in Figure 2 are expressed in lottery tickets. The first figure refers to the payoff to the first mover, the second figure to the payoff to the second mover, and the third payoff to the number of unassigned tickets (details follow later in this section).

Treatment TG is a standard binary trust game, where the outcome of choosing *In* depends on the decision of a second mover. Treatment mTG is a modified binary trust game: while first mover's (he) decisions also affect the payoff of a second player (she), she is passive, and the outcome at the second node is decided by a random draw.[10]

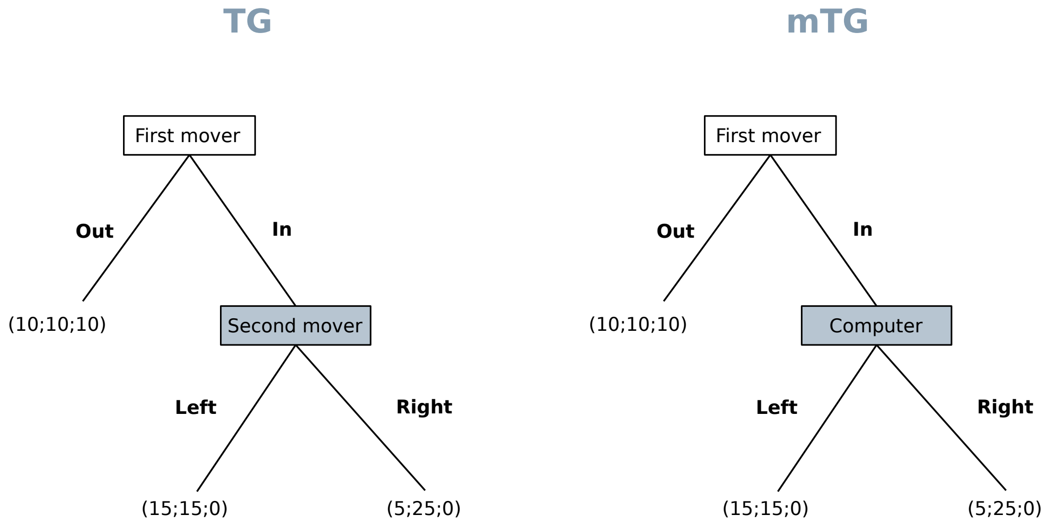

Figure 2: The treatments

Passive second movers (second movers in mTG) did not have to make any decision. Active second movers (second movers in TG) were asked whether they would choose *Left* or *Right*, conditional on their matched first mover choosing *In*. First movers were asked to state a cutoff probability (their minimum acceptable probability, or MAP). In TG, this was: what is the minimum share of opponent decisions that should be *Left* among the decisions made by all potential matches for them to prefer *In* over *Out*? In mTG, first movers were told that other participants play TG. They were also told that the distribution from which the computer makes

[10] mTG is equivalent to BZ's Risky Dictator game. I refer to this treatment as 'mTG' for consistency with the companion paper, Polipciuc and Strobel (2022).

a random draw at the second node is identical with the distribution of second mover decisions in the corresponding TG. They were asked what the minimum share of *Left* options should be in this distribution for them to prefer *In* over *Out*.

I am interested in how social distance to the opponent affects first mover behavior soon after the groups have been formed (at T1) and seven months later (at T2). In each of the two experiments (at T1 and T2), the study combines a between-subject design (each subject is exposed to only one treatment) with a within-subject design (each subject makes decisions for an ingroup and for an outgroup opponent). Importantly, within opponent type (in- or outgroup), the description of how the probability of *Left*, p^*, had been generated is the same in the two treatments. This allows me to investigate how the identity of the opponent affects taking strategically versus randomly generated risks, independently of the effect of beliefs about the trustworthiness of in-/outgroup members.

The social groups I use are groups of about 60 students. All participants are first year students enrolled in the same study track. The groups have been created by the administration office at the beginning of the academic year through random assignment conditional on nationality. Students interact more with members of their own group throughout their first year of study: in all the classes they take, their classmates are from the same group, and they participate in social activities with members of their group only. While they do interact with the rest of their cohort, I assume that the social groups matter enough to create a feeling of in-/outgroup as time passes between T1 and T2.[11]

I chose this group identity for two reasons. First, because it is a natural identity (meaning it has validity outside the lab) which has been assigned randomly. Second, because it falls under what Lane (2016) calls "social/geographical affiliation". Many laboratory studies on discrimination use artificial identities, which are induced during the experiment. The main argument is that this allows for a clean causal identification of a lower bound effect of discrimination: should participants discriminate in this setting with no pre-existing stereotypes—the argument goes—they will discriminate even more outside the lab. However, in his meta-analysis of experimental studies on discrimination, Lane (2016) shows that studies using artificial identities usually report more discrimination than studies using natural identities. This casts doubt on the above-mentioned assumption and provides a reason to study discrimination with natural identities. Among natural identities, Lane (2016) finds that discrimination is most prominent for groups which are divided socially or geographically, such as the student groups used in this study. This suggests it is more likely for measurable discrimination to exist among social/geographical groups.

[11] For details about the social groups and tests of the assumption that students perceived inversus outgroup members differently, see Appendix A.

Table 1: What can be identified if values to ingroup and outgroup differ?

	Determinants of trust	Types of discrimination
TG	(Risk aversion)	NA
	Beliefs about trustworthiness	Statistical discrimination
	Outcome-based social preferences	Outcome-based taste-based discrimination
	Intention-based social preferences	Intention-based taste-based discrimination
mTG	(Risk aversion)	NA
	Beliefs about trustworthiness	Statistical discrimination
	Outcome-based social preferences	Outcome-based taste-based discrimination

Notes: Column 2 lists determinants of trust identified in the literature which manifest in each treatment. I put risk aversion between parentheses, as there is less consensus about it being a determinant of trust. Column 3 states which types of discrimination can be identified if the value of the corresponding determinant differs for in- versus outgroup members. 'NA' stands for 'not applicable'.

Column 2 in Table 1 shows which determinants of trust potentially play a role in each treatment. By contrasting behavior in the two treatments, it is possible to isolate the effect of intention-based social preferences (in this case, betrayal aversion) on trust. Column 3 specifies for each determinant of trust what type of discrimination would ensue if the values of the determinant differ with the social identity of the opponent (in- or outgroup). I argue that ingroup bias (or outgroup favoritism) in betrayal aversion—identified as a difference in differences between behavior in the two treatments and behavior towards the two types of opponents—reflects the part of taste-based discrimination due to intention-based social preferences.

Below I present the timeline of the experiment. There were some procedural differences between T1 and T2 because of different time constraints (a planned limit of 20 minutes at T1 due to external constraints, which was increased to 40 minutes at T2).

1. Upon arrival in the lab, students were asked to which social group they belonged. Within each group, they were then given a code. The code determined the treatment (TG or mTG), the role (first mover or second mover) and the decision order (first movers and active second movers had to make two decisions, one for an outgroup and one for an ingroup opponent). The code also determined who their in- and outgroup opponents would be. Codes were generated such that all role and treatment combinations would be covered within each social group.[12] The experiment ended here for passive second movers. First movers and active second movers received instructions according to their treatment/role combination.

[12] For details about the assignment to treatment and role and about the matching procedure, see Appendix B.

2. First movers and active second movers went through a set of comprehension questions. At T1, they had only one try. If they made mistakes, they received feedback on screen. All participants were allowed to continue to the decision-making part. However, to ensure that I report behavior of participants who understood the instructions, from T1 I include in the estimation sample only those participants who answered the comprehension questions correctly or made a minor mistake.[13] This leads to only one third of first movers at T1 being included in the T1 estimation sample.[14]

At T2, after being able to use only one third of the data at T1, participants could to spend up to 40 minutes in the experiment. If they made mistakes, they received explanations in person from the research team until they answered all comprehension questions correctly. This is why all first movers at T2 are included in the T2 estimation sample.[15]

Table 2 describes the samples at T1 and T2.

3. First movers and active second movers made two decisions, one for an ingroup and one for an outgroup member. They were informed that one of their two decisions will be selected at random to determine their final payoff.

4. First movers and active second movers reported their gender and their risk preferences, positive and negative reciprocity, and generalized trust by answering questions from the survey preference module of Falk et al. (2016). They also had to allocate a lottery ticket (hypothetically) to either an ingroup member or to any participant in the experiment. I use this as a proxy for ingroup favoritism.

Participants were recruited by running the experiments jointly with other experiments for course credit. The two experiments in this study were remunerated separately. As mentioned before, participants in these experiments were paid in lottery tickets. With payment in lottery tickets, it is necessary to have blank tickets to preserve the relative efficiency of outcomes: this is why there are 10 blank tickets if *Out* is implemented, and none if *In* is implemented. Each participant received a total number of tickets equal to his/her final payoff plus a show-up fee of 3 tickets. At both T1 and T2, 15 tickets were drawn after all sessions had taken place and their owners received vouchers worth €100 each. If a blank ticket was drawn, the respective voucher was not awarded. For first movers and active second

[13] I consider a minor mistake to be adding the show-up fee to the correct answer when asked about final payoffs.

[14] Instructions for experiments on betrayal aversion are complex, especially for first movers. Quercia (2016) also finds that at most 40% of first movers answer a similar (but smaller) set of understanding questions correctly from the first try (see his summary statistics of wrong answers to Questions 2 and 3 in the OE (open-ended) elicitation of betrayal aversion, in Table 3, on p. 57).

[15] In Appendix C, I run balancing tests to check whether the samples in the two experiments differ significantly due to this decision. This is not the case for existing observables, such as gender, risk aversion, or negative or positive reciprocity.

Table 2: Participants by treatment at T1 and T2

Experiment	Treatment	Role	# assigned subjects	# subjects with minor/no understanding mistakes	% subjects with minor/no understanding mistakes	# first movers in the estimation sample
T1	TG	First mover	92	41	45%	41
		Second mover	34	20	59%	–
	mTG	First mover	81	24	30%	24
		Second mover	27	–	–	–
T2	TG	First mover	46	31	67%	46
		Second mover	48	44	92%	–
	mTG	First mover	47	19	40%	47
		Second mover	78 [a]	–	–	–

Notes: At T1, I restricted the estimation sample to first movers with minor/no understanding mistakes. At T2, all first movers with valid answers were included in the estimation sample.

[a] At T2 I assigned second mover roles in mTG to participants in another experiment, as their presence in the laboratory at the same time was not necessary. This difference between T1 and T2 should not affect first mover decisions: at both stages, players were informed they had already been matched (in the sense that the matching rule had been decided) and were not told explicitly that their opponent was in the room at the same time.

movers (who were the ones spending more time in the lab), the median duration of the experiment was 13.6 minutes at T1 (10.7 minutes at T2), the maximum duration was 32.4 minutes at T1 (24 minutes at T2), and the chance of winning a voucher was 2% at T1 and 3.3% at T2.[16] The chances were calculated *post factum*. What participants knew was that there were 15 vouchers available, and that there were approximately 700 participants at T1 (600 participants at T2).[17]

The experiments were run in Qualtrics.

4 Conceptual framework and hypotheses

I denote $MAP_{TG1,I}$ as the first movers' MAP in treatment TG with an ingroup opponent at T1, and MAP_{TG1} as the first movers' MAP at T1, regardless of opponent type. The notation is similar for all the other treatment-opponent type combinations: I refers to ingroup matches, O—to outgroup matches.

[16] This translates into a median expected payoff of €8.82 per hour at T1 (€18.5 per hour at T2). The expected payoffs vary between T1 and T2 because there were fewer participants at T2, and because I expected participants to spend more time on average in the laboratory at T2, when in fact the opposite happened. Detailed calculations of the (*a posteriori*) winning chances are available upon request.

[17] These numbers are higher than the totals in Table 2 as there were additional treatments, not discussed in this paper.

The hypotheses fall into two categories: those about behavior at T1 and T2, respectively, and those about the change in behavior between T1 and T2. In the first category, there are hypotheses about discrimination in trust, about the existence of (positive) betrayal aversion, and about discrimination in betrayal aversion at a certain time. In the second category, there are hypotheses about changes in the three concepts between T1 and T2.

4.1 Behavior at T1

Hypothesis 1: $MAP_{TG1,I} = MAP_{TG1,O}$.
Hypothesis 2: $MAP_{mTG1,I} = MAP_{mTG1,O}$.
Hypothesis 3: $MAP_{TG1,O} - MAP_{mTG1,O} = MAP_{TG1,I} - MAP_{mTG1,I} > 0$.

The set of hypotheses at T1 states that I expect to replicate BZ's finding that betrayal aversion exists and is positive, but that I do not expect social group identity to be relevant for trusting decisions at this point (Hypothesis 3). That is, I expect the willingness to accept the risky payoff from trusting ingroup members and that from trusting outgroup members to not differ from each other (Hypothesis 1). I also expect that the identity of the opponent makes no difference for the threshold required to be willing to take the risky bet with payoff externalities for a passive opponent (Hypothesis 2). If Hypothesis 1 and Hypothesis 2 hold simultaneously, then the identity of the opponent also does not affect betrayal aversion at T1.

4.2 Behavior at T2

Hypothesis 4: $MAP_{TG2,I} < MAP_{TG2,O}$.
Hypothesis 5: $MAP_{mTG2,I} < MAP_{mTG2,O}$.
Hypothesis 6: $MAP_{TG2,O} - MAP_{mTG2,O} > MAP_{TG2,I} - MAP_{mTG2,I} > 0$.

The set of hypotheses at T2 draws on Bacine and Eckel's (2018) findings. Despite the fact that MAP_{mTG} and MAP_{TG} cover slightly different concepts from theirs (see Section 3 for details), *a priori* I expect to find the same relationships as they do.

Bacine and Eckel (2018) run their experiment once, a couple of weeks after their subjects were randomly assigned to natural groups. While the timing is more similar to that of the first experiment in this paper, the social identity used in their study is arguably stronger: the residential college in which students live. Because of this, I believe it is more plausible that a weaker identity like the the one used in this study needs a longer time to produce effects. I thus assume the effects found by Bacine and Eckel (2018) are more likely at T2.[18]

[18] The exact timing of T2 was chosen for practical reasons: it had to be towards the end of the

I expect to find a lower MAP_{TG} for ingroup opponents than for outgroup opponents at T2 (Hypothesis 4). This builds on previous experimental findings on unconditional decisions to trust in- versus outgroup members (Lane, 2016).

As Table 1 shows, behavior in mTG reflects risk preferences, beliefs about trustworthiness, and outcome-based social preferences. Risk preferences should not vary with social distance to the opponent generating (part of) the risk. Previous literature suggests that outcome-based social preferences differ towards in- and outgroup members, with individuals being more altruistic towards ingroup members (Li, 2020). Since the time between T1 and T2 is rather short, I assume that at T2 there do not yet exist stereotypes regarding the trustworthiness of members of a specific group. This is why I expect beliefs about the opponent's trustworthiness to either not differ for in- and outgroup members, or to be more optimistic for ingroup members. These effects together lead to Hypothesis 5: I expect first movers to require a lower MAP_{mTG} from an ingroup opponent relative to the one they require from an outgroup opponent.

Finally, Hypothesis 6 means I expect to find betrayal aversion against both in- and outgroup members, with betrayal aversion against outgroup members being higher—similar to the result of Bacine and Eckel (2018).

4.3 Behavior change between T1 and T2

Hypotheses about behavior change are exploratory, as this study is the first—to my knowledge—to measure trust game senders' behavior at two points in time and to use treatments to isolate changes in (discrimination in) betrayal aversion.

Hypotheses in this subsection are a consequence of hypotheses in subsections 4.1 and 4.2 being supported.

Hypothesis 7a: $\Delta MAP_{TG,I} = MAP_{TG2,I} - MAP_{TG1,I} < 0$.

Hypothesis 7b: $\Delta MAP_{TG,O} = MAP_{TG2,O} - MAP_{TG1,O} > 0$.

The hypotheses above refer to changes in TG from T1 to T2. First movers are either more willing to trust ingroup members at T2 than at T1 (H7a) or less willing to trust outgroup members at T2 than at T1 (H7b), or both.

Hypothesis 8a: $\Delta MAP_{mTG,I} = MAP_{mTG2,I} - MAP_{mTG1,I} < 0$.

Hypothesis 8b: $\Delta MAP_{mTG,O} = MAP_{mTG2,O} - MAP_{mTG1,O} > 0$.

Hypotheses 8a and 8b above refer to changes in mTG. First movers are either more willing to enter the modified trust game with ingroup members at T2 than at T1 (H8a) or less willing to enter the modified trust game with outgroup members at T2 than at T1 (H8b), or both.

academic year and to maximize the chance to have a large share of the target student population in the lab. This was possible by recruiting students in compulsory courses, with the help of course coordinators who agreed to this.

Hypothesis 9a: $\Delta BA_I = \Delta MAP_{TG,I} - \Delta MAP_{mTG,I} \leq 0$.
Hypothesis 9b: $\Delta BA_O = \Delta MAP_{TG,O} - \Delta MAP_{mTG,O} \geq 0$.

Hypotheses 9a and 9b above refer to changes in betrayal aversion in time. Between T1 and T2, betrayal aversion towards ingroup members does not increase (H9a), and betrayal aversion towards outgroup members does not decrease (H9b).

From Hypotheses 9a and 9b follows Hypothesis 10:
Hypothesis 10: $\Delta BA_I \leq \Delta BA_O$.

This hypothesis refers to changes in discrimination in betrayal aversion in time. Between T1 and T2, the slope of the change in betrayal aversion towards ingroup members is lower than or equal to the slope of the change in betrayal aversion towards outgroup members.

5 Data and results

5.1 Summary statistics and nonparametric tests

Table 3 presents the summary statistics of MAPs in each treatment at T1 and T2. The upper panel contains data on both decisions. The middle and lower panels report statistics by the opponent's identity. *P*-values in the table are from two-sided Mann-Whitney tests. All *p*-values reported in this section are two-sided.

At T1, there is weak evidence for the existence of betrayal aversion in the pooled sample (*p*-value = 0.071), and also towards ingroup opponents (*p*-value = 0.068). At T2, there is no evidence of betrayal aversion. Between T1 and T2, the MAPs in mTG are the only ones to increase significantly (*p*-value = 0.016, for all MAPs; *p*-value = 0.039, for ingroup matches, both not reported in the table), making betrayal aversion vanish at T2. In the pooled sample of the two decisions with both in- and outgroup matches, MAP_{TG} does not differ between in- and outgroup matches, neither at T1, nor at T2 (*p*-value = 0.833 at T1; *p*-value = 0.285 at T2).

Next, I examine changes in behavior between T1 and T2. MAP_{mTG} at T2 compared to T1 increases in both ingroup and outgroup matches, but the increase between the two periods is significant only in ingroup matches (*p*-value = 0.039 for $MAP_{mTG2,I}$ versus $MAP_{mTG1,I}$; *p*-value = 0.180 for $MAP_{mTG2,O}$ versus $MAP_{mTG1,O}$). MAP_{TG} in both in- and outgroup matches is not significantly different at T1 from T2 (*p*-value = 0.457 for $MAP_{TG2,I}$ versus $MAP_{TG1,I}$; *p*-value = 0.749 for $MAP_{TG2,O}$ versus $MAP_{TG1,O}$).[19]

[19] The comparative statics do not change if I only consider first decisions (not reported).

Table 3: Minimum acceptable probabilities

	In both types of matches			
	TG1	mTG1	TG2	mTG2
	61.15 (18.99)	52.67 (24.03)	57.77 (26.47)	61.51 (22.61)
p-values	0.071		0.437	
Observations	82	48	92	94
	In ingroup matches			
	TG1	mTG1	TG2	mTG2
	61.60 (17.85)	50.38 (24.39)	55.13 (25.73)	60.84 (22.67)
p-values	0.068		0.298	
Observations	41	24	46	47
	In outgroup matches			
	TG1	mTG1	TG2	mTG2
	60.71 (20.29)	54.96 (23.96)	60.41 (27.22)	62.17 (22.78)
p-values	0.443		0.902	
Observations	41	24	46	47
Individuals	41	24	46	47

Notes: 'TG1' refers to TG at T1, 'mTG1' to mTG at T1, etc. The table shows averages per treatment. Each participant made two decisions. P-values are from ranksum tests of behavior with the two opponent types (active in TG, or passive in mTG) in an experiment (at T1 or T2). Standard deviations in parentheses.

5.2 Behavior at T1 and T2

The tests reported above do not take into account that the same participants make two decisions (for an in- and for an outgroup opponent). To account for this, I run regression analyses separately for the samples at T1 (in Table 4) and T2 (in Table 5). In these regressions, I cluster errors at the individual level.

Table 4: Linear regressions on Minimum Acceptable Probabilities at T1

	(1)	(2)	(3)	(4)
Baseline: mTG1, ingroup match				
TG1	10.58*	8.92	15.17**	14.98**
	(6.01)	(5.76)	(5.71)	(5.61)
Outgroup match	4.58	4.58	3.18	3.18
	(3.69)	(3.72)	(4.08)	(4.13)
TG1 × Outgroup match	−5.48	−5.48	−2.37	−2.37
	(4.98)	(5.02)	(5.25)	(5.31)
Risk loving (0–10)		−2.44*		−1.65
		(1.24)		(1.36)
Male		3.95		−0.59
		(4.29)		(5.47)
Ingroup first	2.74	1.23	−6.04	−7.61
	(4.87)	(4.79)	(5.27)	(5.76)
Constant	49.35***	67.41***	36.19***	46.24***
	(5.19)	(10.09)	(8.66)	(11.86)
Session fixed effects	No	No	Yes	Yes
Adjusted R^2	0.02	0.04	0.16	0.16
Observations	130	130	104	104
Individuals	65	65	52	52
Sessions	33	33	20	20

Notes: Standard errors clustered at the individual level in parentheses. 'TG1' refers to TG at T1, 'mTG1' to mTG at T1, etc. The sample in models (1) and (2) consists of first movers with minor/no understanding mistakes. The sample in models (3) and (4) consists of those first movers with minor/no understanding mistakes who were not the only ones in their session to fulfill this criterion. One can only add session fixed effects for this smaller second sample. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In Table 4, the dependent variable is the MAP. Standard errors are clustered at the individual level. Models (1) and (2) do not include session fixed effects, while models (3) and (4) do. As I move from column (1) to (2), and from (3) to (4), I add the control variables mentioned in the table. The baseline is $MAP_{mTG1,I}$. The coefficient for TG1 (which reflects betrayal aversion towards ingroup members at T1) is positive in all four specifications, with the coefficients in last two columns being significant at p-value < 0.05. Playing with an in- as opposed to an outgroup opponent does not make a difference at T1 in either of the two treatments (from Wald tests for equality of coefficients: in (4), $MAP_{mTG1,I}$ versus $MAP_{mTG1,O}$: p-value $= 0.445$; $MAP_{TG1,I}$ versus $MAP_{TG1,O}$: p-value $= 0.808$). Results in columns (3) and (4) however should be taken as an indication: to include session fixed effects, I had to restrict the sample to those first movers with minor or no understanding mistakes who were not the only ones in their session to fulfill this requirement. This leaves a small sample scattered across a considerable number of sessions.[20,21]

Result 1. At T1, I find evidence of betrayal aversion. There is no discrimination in betrayal aversion towards second movers from the in- versus the outgroup. There is also no difference in the willingness to accept the risky payoff from trusting in- versus outgroup members.

Result 1 largely supports Hypotheses 1, 2, and 3.

[20] I nonetheless report results in columns (3) and (4), as results from the companion paper, Polipciuc and Strobel (2022)—which includes additional treatments and thus has a bigger sample—confirm the sign and the significance level of betrayal aversion at T1.

[21] In the full sample, the coefficients of TG1, Outgroup match and their interaction have the same signs as in Table 4, but are insignificant. Results available on request.

Table 5: Linear regressions on Minimum Acceptable Probabilities at T2

	(1)	(2)	(3)	(4)
Baseline: mTG2, ingroup match				
TG2	−5.76	−6.30	−6.95	−7.77
	(5.07)	(5.10)	(4.87)	(4.92)
Outgroup match	1.33	1.14	1.33	1.14
	(2.13)	(2.18)	(2.19)	(2.24)
TG2 × Outgroup match	3.96	4.15	3.96	4.15
	(3.56)	(3.61)	(3.66)	(3.70)
Risk loving (0–10)		−3.20***		−3.01**
		(1.11)		(1.19)
Male		−0.17		−2.10
		(5.20)		(5.17)
Ingroup first	−4.00	−6.05	−7.06	−8.29*
	(4.81)	(4.61)	(4.70)	(4.50)
Constant	62.80***	86.88***	59.10***	78.02***
	(4.20)	(8.68)	(6.93)	(9.41)
Session fixed effects	No	No	Yes	Yes
Adjusted R^2	−0.00	0.06	0.08	0.14
Observations	186	184	186	184
Individuals	93	92	93	92
Sessions	10	10	10	10

Notes: Standard errors clustered at the individual level in parentheses. The sample in models (2) and (4) has one respondent fewer than the one in models (1) and (3), because one participant did not specify their gender. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In Table 5, I repeat the same exercise for the data collected at T2. Here, in all four specifications, the coefficient quantifying betrayal aversion towards the ingroup is negative and insignificant. A Wald test of equality of coefficients in column (4) indicates that $MAP_{TG2,I}$ is marginally lower than $MAP_{TG2,O}$ (p-value $= 0.077$).

Since the coefficient for playing against an ingroup opponent first is also weakly significant in specification (4), I check whether there are order of play effects (playing first with an in- or an outgroup opponent). I do this by using a triple interaction, between the treatment dummy, a dummy for facing an outgroup opponent, and one which is 0 for the first decision and 1 for the second decision (see Appendix Table **??**). The test for equality of coefficients shows that when it comes to first decisions only, $MAP_{TG2,I}$ is significantly lower than $MAP_{TG2,O}$ (p-value $= 0.029$). This is supporting evidence that the difference observed in column (4) of Table 5—a marginally lower $MAP_{TG2,I}$ than $MAP_{TG2,O}$, regardless of decision order—is likely not an artifact of decision making order (for instance, due to respondents potentially wishing to be perceived as consistent by the experimenter) as it stems from the cleaner first decision participants make.

Result 2. At T2, I do not find evidence of betrayal aversion, neither towards in-, nor towards outgroup members. For the first of the two decisions, first movers in the TG treatment are more likely to set a more lenient threshold for entering a trusting relationship with an ingroup member than with an outgroup member.

Result 2 does not provide support for Hypotheses 5 and 6. The first of the two decisions first movers make is in line with previous findings on discrimination in trusting members of in- versus outgroups (for unconditional trust). This provides partial support for Hypothesis 4.

For the first decision at T2, the ordering of the MAPs is the following: $MAP_{TG2,I} < MAP_{mTG2,I} < MAP_{mTG2,O} < MAP_{TG2,O}$. For this decision, the direction is that of a strategic risk discount for ingroup matches on average (the opposite of betrayal aversion, so preferring the trusting interaction to the equally risky bet with payoff externalities for another passive participant) and a strategic risk premium for outgroup matches on average (betrayal aversion). However, since the only significant difference is between the first and the last terms in this list of inequalities, I cannot quantify the contribution of each intermediary difference to the difference between the most extreme terms.[22]

Result 3. At T2, the ingroup bias in trust in the first decision cannot be decomposed into a part due to an ingroup bias in betrayal aversion and a residual bias.

Additionally to the hypotheses specified in Section 4, I look into heterogeneous effects at T2 by the strength of the attachment to the ingroup as proxied by the

[22] For the second decision, none of the four MAPs is significantly different from the rest.

hypothetical allocation task. In Appendix Table D.2, I focus on first decisions only. I regress the MAP on the interaction of the treatment dummy, the opponent identity dummy and the dummy for selecting an ingroup recipient in the hypothetical allocation task. In specification (4), I notice that those who select an ingroup recipient in the hypothetical allocation task have a significantly lower $MAP_{TG2,I}$ than $MAP_{TG2,O}$ (p-value = 0.02) and a significantly lower $MAP_{mTG2,I}$ than $MAP_{mTG2,O}$ (p-value = 0.02). Neither of the two is the case for those who select a random recipient from the entire subject pool in the allocation task. For them, neither MAP_{TG2} nor MAP_{mTG2} differ significantly for in- as opposed to outgroup opponents (p-value = 0.45 in TG; p-value = 0.11 in mTG).

I do not find evidence of betrayal aversion for neither first movers who allocate the ticket to an ingroup member nor for first movers who allocate it to a random participant. Since the hypothetical allocation task is a proxy for higher altruism towards the ingroup relative to the outgroup, I interpret these results as evidence that between-subject heterogeneity in outcome-based social preferences is an important factor in explaining the in-/outgroup gap in trust in the first decision at T2. Since variation in behavior in the hypothetical allocation task is endogenous, this evidence is correlational.[23]

Result 4. In first decisions at T2, first movers who give the hypothetical lottery ticket to an ingroup recipient ask for higher MAPs in outgroup matches compared to ingroup matches in both treatments. First movers who give it to a random recipient from the entire subject pool do not state different MAPs in inversus outgroup matches. Neither of the two types of first movers displays betrayal aversion on average, neither towards in- nor towards outgroup opponents.

5.3 Change in behavior between T1 and T2

Ideally, pooled data from the experiments at T1 and T2 would have been in panel format. For privacy reasons, respondents could not be traced between the two periods—but it is highly likely that some respondents at T1 are also present at T2, as both samples are subsets of the same study cohort. This affects the standard errors of regressions on the pooled dataset and might result in different significance levels for some coefficients.

To address this, in Appendix E I use Monte Carlo simulations to estimate how much overlap between the samples can be expected. Then, I check by how much the precision of the estimates of interest can be expected to decrease due to this expected overlap. Results suggest that the expected decrease in precision is small:

[23] In the pooled data on both decisions at T2, only $MAP_{TG2,I}$ is significantly different (lower) than $MAP_{TG2,O}$ among those selecting an ingroup recipient in the hypothetical allocation task (p-value = 0.05). Results available on request.

the 95% confidence intervals for p-values for tests of hypotheses 7–10 over 10,000 simulations span less than 10^{-4}.

This means that the simple regressions presented in Table 6—where the possible overlap in samples at T1 and T2 is unaccounted for—are informative for hypotheses 7–10. In model (1), I regress the MAP on dummies for game type (mTG = 0, TG = 1), opponent type (ingroup = 0, outgroup = 1), and experiment (T1 = 0, T2 = 1), as well as their interactions. In model (2), I add control variables for risk attitudes, gender and a dummy for whether the decision about an ingroup member came first (no = 0, yes = 1). In both models, standard errors are clustered at the individual level, which can only be observed within an experiment (either at T1 or at T2). Models (3) and (4) correspond to (1) and (2), respectively, but the sample at T1 is reduced to those first movers with minor/no understanding mistakes who were not the only ones in their session to fulfill this requirement. Models (3) and (4) include session fixed effects.

Wald tests for equality of coefficients suggest that there is a significant reduction in betrayal aversion towards ingroup members between T1 and T2 (p-value = 0.04 in (2); p-value < 0.01 in (4)). Models without session fixed effects do not allow for more in depth conclusions. In model (4), there is also support for a decrease in betrayal aversion towards outgroup members between T1 and T2 (p-value = 0.03), and marginally significant support for first movers asking for lower MAPs in the modified trust game with ingroup members at T2 than at T1 (p-value = 0.09).

Result 5. Betrayal aversion towards ingroup members decreases significantly between T1 and T2, and so does that towards outgroup members. This is not due to a change in $MAP_{TG,I}$, but to a (marginally significant) increase in $MAP_{mTG,I}$.

In conclusion, there is no support for either part of Hypothesis 7, Hypothesis 8b, or Hypothesis 10. Hypothesis 8a is contradicted (marginally), Hypothesis 9a is supported, and Hypothesis 9b is contradicted.

Table 6: Linear regressions on Minimum Acceptable Probabilities in the pooled data set

	(1)	(2)	(3)	(4)
Baseline: mTG1, ingroup match				
TG	11.23**	9.56*	13.31**	14.78***
	(5.68)	(5.59)	(5.20)	(5.14)
Outgroup match	4.58	4.58	3.18	3.18
	(3.65)	(3.67)	(3.81)	(3.84)
TG × Outgroup match	−5.48	−5.48	−2.37	−2.37
	(4.94)	(4.96)	(4.90)	(4.93)
T2	10.47*	8.78	15.71	24.04*
	(5.95)	(5.78)	(13.12)	(13.87)
TG × T2	−16.94**	−15.73**	−20.03***	−22.52***
	(7.60)	(7.56)	(7.22)	(7.15)
Outgroup match × T2	−3.26	−3.45	−1.86	−2.04
	(4.23)	(4.27)	(4.42)	(4.47)
TG × Outgroup match × T2	9.43	9.62	6.33	6.52
	(6.08)	(6.13)	(6.17)	(6.23)
Risk loving (0–10)		−2.92***		−2.68***
		(0.85)		(0.99)
Male		1.67		−1.56
		(3.55)		(4.25)
Ingroup first		−3.15		−8.30**
		(3.36)		(3.63)
Constant	50.37***	73.92***	35.03***	52.45***
	(4.94)	(7.51)	(9.94)	(11.15)
Session fixed effects			✓	✓
Adjusted R^2	0.00	0.05	0.07	0.14
Observations	316	314	290	288
Individuals	158	157	145	144

Notes: Standard errors clustered at the individual level in parentheses. ‘TG1’ refers to TG at T1, ‘mTG1’ to mTG at T1, etc. The sample in (1) and (2) consists of first movers with minor/no understanding mistakes at T1, and all first movers at T2. The sample in (3) and (4) consists of first movers with minor/no understanding mistakes at T1 who were not the only ones in their session, and all first movers at T2. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. One participant did not specify their gender, which explains the lower number of observations for (2) and (4).

6 Discussion

In this paper, I study experimentally how trust and betrayal aversion vary with social distance, and how the contribution of betrayal aversion to trust changes as social groups develop a group identity. For this purpose, I adapted the design of Aimone and Houser (2012) for identifying betrayal aversion, a concept introduced by Bohnet and Zeckhauser (2004).

I was motivated by evidence of discrimination in trust (Lane, 2016) and by a recently growing literature on the valuation of intentions (Stanca et al., 2009; Strassmair, 2009; Gul and Pesendorfer, 2016; Johnsen and Kvaløy, 2016; Chao, 2018; Orhun, 2018). Several recent papers (Butler and Miller, 2018; Polipciuc and Strobel, 2022) support BZ's interpretation that in a two-player binary trust game, betrayal aversion is the first mover's response to how he perceives the second mover's intentions towards him. That is, betrayal aversion is the result of the first mover preemptively shielding himself from the disutility of a potential betrayal. However, most papers finding evidence of betrayal aversion use an experimental design which does not rule out confounding explanations such as ambiguity aversion, should participants not be rational expected utility maximizers (Li et al., 2020). The design in this study controls for participants' subjective beliefs and measures betrayal aversion net of the confounds listed in Li et al. (2020).

I examine the willingness to accept the risky payoff from trusting in- versus outgroup opponents and its components in a student population at two points in time. Participants have been quasi-randomly assigned to social groups independently of this study. The first experiment takes place shortly after the social groups have been created and the second one seven months later. In the first experiment, betrayal aversion is positive and indiscriminate towards in- and outgroup members. In the second experiment, betrayal aversion to both in- and outgroup members vanishes. In the first of the two decisions they make in the second experiment, first movers set a lower requirement to be willing to trust ingroup members than outgroup members.

When looking more closely, discrimination in trust in the second experiment is a composite of two types of behavior: that of a slight majority (60%) who select a random ingroup recipient over a random recipient from the entire subject pool in a hypothetical allocation task, and the rest, who select an entirely random recipient. Neither of the two groups displays betrayal aversion, neither to in- nor to outgroup opponents. Those who select a completely random recipient do not show intergroup discrimination in trust. The small majority who select an ingroup recipient in the allocation task ask for a significantly higher guarantee of a favorable outcome to trust an outgroup relative to an ingroup member.

The source of risk—random or strategic—is irrelevant in the second experiment. Social distance matters for trust, but only for a small majority of first movers. For this small majority, social distance affects a component which captures risk aversion, beliefs about trustworthiness and outcome-based social preferences towards an opponent.

From these results, I conclude that risk preferences, beliefs about trustworthiness and outcome-based social preferences towards an opponent seem to drive the threshold required to trust or to enter the lottery in the second experiment. In neither of the two experiments is there discrimination in betrayal aversion. While this is at odds with results in the early literature on betrayal aversion, it indirectly supports claims that what had been called betrayal aversion in this early literature might have been confounded, for instance, by ambiguity aversion. Future research should check whether betrayal aversion survives controlling for subjective beliefs.

However, the lack of betrayal aversion in the second experiment could be due to concurrent changes between T1 and T2. Since participants had more time to make a decision at T2 (up to 40 minutes versus 20 minutes at T1), they were more likely to make more analytical, "System 2" (Kahneman, 2003) type of decisions at this time. It is also possible that the effect is at least partially explained by selection: students who passed their exams (including two statistics exams) are more likely to be part of the study program by T2. This means the sample at T2 might be more analytical on average than the sample at T1, and less prone to emotional reactions such as betrayal aversion (Aimone and Houser, 2011; Aimone et al., 2015).

Finally, I note the characteristics of the setup in which this null result was found, to facilitate the comparison with related studies: (i) the social groups were formed outside the laboratory, in a type of setting which Lane (2016) found to be the most conducive to intergroup discrimination; (ii) group identity was assigned randomly, making causal inferences about the effect of group identity cleaner (at the minimum, in the first experiment—there is attrition between T1 and T2, as students who drop out are not anymore in the sample at T2); (iv) participants are business and economics students in a developed country; (v) the task is highly stylized; and (vi) the social identity used does not entail a conflict over resources, nor competition between groups outside the laboratory.

References

Aimone, J. A., Ball, S., and King-Casas, B. (2015). The betrayal aversion elicitation task: An individual level betrayal aversion measure. *PLoS ONE*, 10(9):e0137491.

Aimone, J. A. and Houser, D. (2011). Beneficial betrayal aversion. *PLoS ONE*, 6(3):e17725.

Aimone, J. A. and Houser, D. (2012). What you don't know won't hurt you: A laboratory analysis of betrayal aversion. *Experimental Economics*, 15(4):571–588.

Arrow, K. J. (1973). The theory of discrimination. In Ashenfelter, O. and Rees, A., editors, *Discrimination in labor markets*, pages 3–33. Princeton, NJ: Princeton University Press.

Arrow, K. J. (1974). *Limits of Organization.* W. W. Norton & Company.

Ashraf, N., Bohnet, I., and Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics*, 9(3):193–208.

Bacine, N. and Eckel, C. C. (2018). Trust and betrayal: An investigation into the influence of identity. Working paper.

Balliet, D., Wu, J., and Dreu, C. K. W. D. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, 140(6):1556–1581.

Becker, G. M., De Groot, M. H., and Marshak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9:226–232.

Becker, G. S. (2010). *The economics of discrimination.* University of Chicago press.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior*, 10(1):122–142.

Binzel, C. and Fehr, D. (2013). Social distance and trust: Experimental evidence from a slum in Cairo. *Journal of Development Economics*, 103:99–106.

Bohnet, I., Greig, F., Herrmann, B., and Zeckhauser, R. (2008). Betrayal aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, 98(1):294–310.

Bohnet, I., Herrmann, B., and Zeckhauser, R. (2010). Trust and the reference points for trustworthiness in Gulf and Western countries. *Quarterly Journal of Economics*, 125(2):811–828.

Bohnet, I. and Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4):467–484.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. (2019). Inaccurate statistical discrimination: An identification problem. NBER Working Paper No. 25935.

Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398.

Breuer, W. and Hüwe, A. (2014). Trust, reciprocity, and betrayal aversion: Theoretical and experimental insights. Working paper.

Butler, J. V., Giuliano, P., and Guiso, L. (2016). The right amount of trust. *Journal of the European Economic Association*, 14(5):1155–1180.

Butler, J. V. and Miller, J. B. (2018). Social risk and the dimensionality of intentions. *Management Science*, 64(6):2787–2796.

Chao, M. (2018). Intentions-based reciprocity to monetary and non-monetary gifts. *Games*, 9(4):74.

Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869.

Chuah, S.-H., Fahoum, R., and Hoffmann, R. (2013). Fractionalization and trust in India: A field-experiment. *Economics Letters*, 119(2):191–194.

Costa-Gomes, M. A., Huck, S., and Weizsäcker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, 88:298–309.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, 46(2):260 – 281.

Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, 59(1):17–45.

Cox, J. C., Friedman, D., and Sadiraj, V. (2008). Revealed altruism. *Econometrica*, 76(1):31–69.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.

Engelmann, D., Friedrichsen, J., van Veldhuizen, R., Vorjohann, P., and Winter, J. (2021). Decomposing trust. Personal communication.

Etang, A., Fielding, D., and Knowles, S. (2010). Does trust extend beyond the village? Experimental trust and social distance in Cameroon. *Experimental Economics*, 14(1):15–35.

Fairley, K., Sanfey, A., Vyrastekova, J., and Weitzel, U. (2016). Trust and risk revisited. *Journal of Economic Psychology*, 57:74–85.

Falk, A., Becker, A., Dohmen, T., Huffman, D. B., and Sunde, U. (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences. IZA Discussion Paper No. 9674.

Falk, A. and Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, 54(2):293–315.

Falk, A. and Zehnder, C. (2013). A city-wide experiment on trust discrimination. *Journal of Public Economics*, 100:15–27.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the European Economic Association*, 7(2-3):235–266.

Fershtman, C. and Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *The Quarterly Journal of Economics*, 116(1):351–377.

Fetchenhauer, D. and Dunning, D. (2012). Betrayal aversion versus principled trustfulness—how to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization*, 81(2):534–541.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., and Soutter, C. L. (2000). Measuring trust. *Quarterly Journal of Economics*, 115(3):811–846.

Guillen, P. and Ji, D. (2011). Trust, discrimination and acculturation. *The Journal of Socio-Economics*, 40(5):594–608.

Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural biases in economic exchange? *Quarterly Journal of Economics*, 124(3):1095–1131.

Gul, F. and Pesendorfer, W. (2016). Interdependent preference models as a theory of intentions. *Journal of Economic Theory*, 165:179–208.

Güth, W., Levati, M. V., and Ploner, M. (2008). Social identity and trust—an experimental investigation. *The Journal of Socio-Economics*, 37(4):1293–1308.

Hargreaves Heap, S. P. and Zizzo, D. J. (2009). The value of groups. *American Economic Review*, 99(1):295–323.

Hong, K. and Bohnet, I. (2007). Status and distrust: The relevance of inequality and betrayal aversion. *Journal of Economic Psychology*, 28(2):197–213.

Johnsen, Å. A. and Kvaløy, O. (2016). Does strategic kindness crowd out prosocial behavior? *Journal of Economic Behavior & Organization*, 132:1–11.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5):1449–1475.

La Porta, R., Lopez-de Silanes, F., Schleifer, A., and Vishny, R. W. (1997). Trust in large organizations. *The American Economic Review.*

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review*, 90:375–402.

Li, C., Turmunkh, U., and Wakker, P. P. (2020). Social and strategic ambiguity versus betrayal aversion. *Games and Economic Behavior*, 123:272–287.

Li, S. X. (2020). Group identity, ingroup favoritism, and discrimination. In Zimmermann, K., editor, *Handbook of Labor, Human Resources and Population Economics*. Springer International Publishing.

Orhun, A. Y. (2018). Perceived motives and reciprocity. *Games and Economic Behavior*, 109:436–451.

Polipciuc, M. and Strobel, M. (2022). Betrayal aversion with and without a motive. Working paper.

Quercia, S. (2016). Eliciting and measuring betrayal aversion using the BDM mechanism. *Journal of the Economic Science Association*, 2(1):48–59.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):1281–1302.

Sapienza, P., Toldra-Simats, A., and Zingales, L. (2013). Understanding trust. *The Economic Journal*, 123(573):1313–1332.

Schwerter, F. and Zimmermann, F. (2020). Determinants of trust: The role of personal experiences. *Games and Economic Behavior*, 122:413–425.

Stanca, L., Bruni, L., and Corazzini, L. (2009). Testing theories of reciprocity: Do motivations matter? *Journal of Economic Behavior & Organization*, 71(2):233–245.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, 38(2):332–382.

Strassmair, C. (2009). Can intentions spoil the kindness of a gift? An experimental study. Working paper, University of Munich.

Suchon, R. and Villeval, M. C. (2019). The effects of status mobility and group identity on trust. *Journal of Economic Behavior & Organization*, 163:430–463.

Zak, P. J. and Knack, S. (2001). Trust and growth. *The Economic Journal*, 111(470):295–321.

Appendix A The social groups: creating the in-group/outgroup

Independently from this study, the administration office worked together with an important student association to create so-called *communities* for first-year bachelor students. The communities' purpose was to "create social bonds, build friendships and study together" (personal communication with the administration office).

In years before the experiments were run, students would attend tutorials (in groups of about 15 students) with a random selection of classmates from the entire cohort enrolled in the study track (approximately 640 students in 2017/2018) in each course. In 2017/2018, the pool from which their classmates were selected was reduced to their community. Thus, the pool became approximately ten times smaller (the mean community size was 62 students). This led to students spending more time with members of their own community in class and preparing for class.[24]

Moreover, weekly social meetings were organized for each community. A couple of second-year volunteer students (student guides) were assigned to each community to answer study-related questions and to help organize social activities (dinners, trips, film evenings, sports competitions etc.) for community members. A small budget was allocated by the faculty to support these activities.

I make the assumption that a meaningful distinction was created between in- and outgroup. I check this assumption in two ways.

First, I look at administrative data such as evaluations of the functioning of the tutorial groups (a subset of the community) and of the communities. Unfortunately, students were asked to evaluate the functioning of their community only once, approximately two months after the start of the academic year, and student guides also once, in the middle of the academic year. This means there is no administrative data available to check whether communities became more important later in the academic year relative to a baseline in the beginning. Even so, I report summary statistics of students' evaluations of their tutorial group functioning and of their sense of belonging to their community in Table A.1.

After the first two months of studying students were asked to evaluate their community's functioning.[25] A higher number indicates more agreement with a

[24] Group work is often explicitly required for courses at Maastricht University, e.g. students have to meet to work on a paper which they hand in as a group. Six in ten first-year course descriptions mention 'group work' as a teaching method in this study track. Course descriptions are available at `http://code.unimaas.nl/`, by selecting the bachelor courses for the academic year 2017/2018, and then, for the Bachelor International Business Courses, 'Year 1 Compulsory Courses' and 'Year 1 Compulsory Skills'. Accessed on May 13, 2019.

[25] This is part of the standard course evaluation forms. The data was collected independently of the two experiments in this paper.

statement. Statement 3 is the closest proxy to community attachment. Answers to this question offer moderate support to the assumption that communities created a meaningful in-/outgroup distinction. Students evaluated their communities' and tutorial group functioning positively (but there are no counterfactual or baseline evaluations).

Table A.1: First-year students' assessment of the communities' functioning

Statements	Respondents	Mean	SD	Median
1. The Community program helped me feel like I belong to SBE. (1–5)	598	3.1	1.1	3
2. The SBE Community program helped me to get off to a good start. (1–5)	599	3.2	1.1	3
3. I feel like I belong to the SBE community. (1–5)	602	3.7	1	4

Notes: A higher number indicates more agreement with the statement. 'SBE' stands for the School of Business and Economics.

Second, I included a hypothetical allocation task at the end of both experiments to proxy for ingroup favoritism. The question was: 'Assume you can give one extra lottery ticket to someone else. Who would you give it to?'. The answer options were 'A randomly chosen person from your community' and 'A randomly chosen person who is taking the *[name of the course in which students were recruited]* course'. 54% of first movers select an ingroup member as the preferred recipient at T1, and 60% do so at T2. A two-tailed Mann-Whitney test shows this difference is not significant (p-value = 0.427).

I also use chi-square tests on frequencies, to check whether the answers at T1 (T2) differ significantly from the uniform distribution of 50–50. At T1, the difference is not significant (the p-value for the Pearson chi-square statistic is 0.535). However, at T2 this difference is significant at 5% (p-value = 0.049).[26]

Taken together, these results offer moderate support for the assumption that a meaningful sense of in-/outgroup had developed between T1 and T2.

[26] The results of chi-square tests on intergroup discrimination are similar for the sample of active second movers (p-value = 0.273 at T1—I only consider second movers with correct answers to the understanding questions; p-value = 0.035 at T2). In the combined sample of first movers and active second movers, 63% select a random ingroup recipient at T2. In this combined sample, the frequencies at T2 differ significantly from 50–50 (p-value = 0.002 at T2), while they do not at T1 (from T1, only including those with correct answers: p-value = 0.259). This suggests that ingroup favoritism in altruism may have developed by T2.

Appendix B Assignment to treatment and matching procedure

Note: *This section is very similar to the one described in Appendix B of the companion paper, Polipciuc and Strobel (2022). The reason is that the data collected at T1 for this paper is a subset of the data on which the companion paper is based. This is why the matching procedure is the same.*

The matching procedure ensures that there is a sufficient number of participants in both roles in each treatment from each social group, such that both in- and outgroup matches can be formed truthfully. Participants registered for their preferred time slot online, on a first-come, first-served basis. As a result, social groups were spread unevenly across experimental sessions.

At T1, I assigned the first four individuals in a social group in show-up order to passive second mover roles. At T2, passive second mover roles were assigned to individuals from the same student pool who took part in another experiment. The remaining assignment rules are identical at T1 and T2.

The next (first, at T2) six participants were assigned active second mover roles. Those who arrived to the laboratory after that were assigned in round-robin fashion within each community to first mover roles.[27] The matching was implemented after all data had been collected, according to a matching rule decided upon in advance.

The matching procedure ensured that:

– first movers knew that one of the decisions they made, drawn at random, may also affect the payoff of *another participant*;

– active second movers knew that one of their decisions, drawn at random, may affect the payoff of *other participants*;[28]

– passive second movers knew their payoff was determined either by a computer draw, or jointly by a computer draw and another participant's decision. This was the case for passive second movers at both T1 and T2.

Participants received sheets with unique randomly generated four-digit codes. These sheets accompanied the instructions. Within each treatment-role-opponent type pool of subjects (in-/outgroup) across participants in all sessions in an experiment (T1 or T2), participants were sorted by this code. After the random draw which decided whether the in- or the outgroup decision was selected for a participant, matches were created by assigning the first first mover in a pool to the

[27] The round-robin assignment to treatment also alternated the order in which participants made the two decisions, for an out- and for an ingroup member, respectively.

[28] I chose for this asymmetry in the instructions for first movers and active second movers because I was interested solely in first mover decisions. I thus wanted to maximize the number of subjects assigned to first mover roles, while still being able to truthfully create in- and outgroup matches for all participants. This meant that a second mover would be matched to multiple first movers.

first second mover in the corresponding pool, the second first mover to the second second mover, etc. Participants were aware that they had already been matched when making their decisions. This is true in the sense that the matching rule had already been set.[29]

A couple of weeks after the data collection for the respective experiment ended, 15 lottery tickets were drawn at random from all tickets. The winners had to present the sheet with the winning lottery code to a third party not involved in running the experiment to collect their earnings.[30]

[29] As Butler and Miller (2018) mention, matching prior to decision making is important: first movers know that when they have an active opponent, if *In* is implemented, they get her decision, rather than a decision drawn from a pool of decisions. This makes the difference with having a passive opponent more salient.

[30] There is a slight difference in the way I computed payoffs for matches in TG versus mTG if the MAP was greater than or equal to p^*. This difference is the same at both T1 and T2. In TG, first movers each received a second mover's choice—akin to a draw *without replacement* from a pool of decisions. In mTG, all first movers got a draw from the same urn, *with replacement.* While this does not influence one's own chance of receiving a certain payoff, it does affect the outcome distributions in the two treatments.

I became aware of this difference *post factum.* The difference should not have affected first mover decisions in TG and mTG, as it was not apparent in the instructions. The instructions only described how one's own p^* is calculated, without any reference to the chances faced by other first movers.

Appendix C Balancing tests and robustness checks

A balancing test in Appendix Table C.1 shows that active participants were similarly likely to answer the five understanding questions common to both treatments (TG and mTG) correctly in both treatments at T1. First movers in mTG had an additional understanding question, about how the probability distribution of draws they faced was linked to actual choices of active second movers in the corresponding (in-/outgroup) TG.

Table C.1: Predictors of answering the five understanding questions common to both treatments with minor/no mistakes at T1

TG1	0.12
	(0.14)
Time spent reading instructions (min)	0.06***
	(0.01)
TG1 × Time spent reading instructions (min)	0.00
	(0.03)
Risk loving (0–10)	−0.01
	(0.02)
Male	−0.14
	(0.10)
Constant	0.28
	(0.18)
Adjusted R^2	0.13
Individuals	173

Notes: The estimation sample includes all first movers at T1. The baseline is mTG. I interacted time spent reading instructions with facing an active opponent because the word count differs in the two situations (passive second mover: 1,021 words; active second mover: 911 words). Standard errors are clustered at session level. Regressions include session fixed effects. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

I also run balancing tests on observable characteristics in Appendix Table C.2, to check whether the estimation samples at T1 and T2 differ significantly on any of these characteristics. The answers to the five understanding questions common to both treatments are more likely to be correct from the first try at T2 than at T1 in the full sample. Participants at T1 were less likely to answer the understanding questions correctly from the first try.

By construction, the estimation sample only includes those respondents from T1 who answered the understanding correctly, but all participants with valid answers at T2. This is why in the estimation sample participants at T1 are more likely than participants at T2 to answer the common understanding questions correctly from the first try. Since the share of correct answers to the five common understanding questions was lower in mTG than in TG at T1 (30% versus 45%, see Table 2), the estimation sample at T1 has more first movers in TG than in mTG. The shares of first movers in TG and mTG are balanced at T2—as a result, it is more likely that if a participant in mTG is present in the estimation sample, this is the case at T2.

This could potentially pose a problem if the MAPs in mTG at T1 differ substantially among those who answer the understanding questions correctly and those who do not. Two-tailed Mann-Whitney ranksum tests show that the MAPs do not differ significantly between these groups (p-value $= 0.16$ for both decisions, p-value $= 0.14$ for the first decision only).

In both types of samples (the full samples and the estimation samples), the time spent reading the instructions is shorter at T2, possibly due to a better command of English at T2 than at T1.[31] While none of the individual characteristics are unbalanced, I do notice that in the estimation samples, dropping those with incorrect answers at T1 led to fewer individuals being assigned to mTG at T1.

This raises the question of how selection affects the generalizability of the results. To check this, I examine the behavior of respondents at T2 who have not answered the comprehension questions correctly from the first try. Since at T2 these individuals received comprehensive feedback, I assume that by the time they report their MAPs, they had understood the instructions, just as those who had answered these questions correctly from the first try. Should the answers at T2 of these two types of first movers—those who answered correctly or made a minor mistake versus those who answered incorrectly—differ substantially, this would be reason to believe that by dropping those with incorrect answers at T1, I may have dropped a certain type of responses. This is however not the case: only MAPs in mTG are marginally higher for those who answered correctly (p-value < 0.1).

[31] Experiments at the Behavioral and Experimental Economics Laboratory at Maastricht University are carried out in English, which is the language of instruction for students in the target population. However, most students' first language is not English—so it is likely that their level of English improves considerably in their first year of studies.

Table C.2: Balancing tests: do samples at T1 and T2 differ?

	Full samples T1 + T2	Estimation samples T1 + T2
Comprehension questions answered correctly from the first try	0.208 *** (0.066)	−0.387 *** (0.056)
Time spent reading instructions (min)	−0.805 ** (0.342)	−1.664 *** (0.432)
Total duration (min)	−1.047 * (0.625)	−0.634 (0.663)
Was in mTG	0.037 (0.037)	0.136 ** (0.059)
Male	−0.083 (0.063)	−0.070 (0.082)
Risk loving (0–10)	−0.293 (0.274)	−0.271 (0.325)
Others can be trusted	0.073 (0.295)	0.093 (0.330)
Positively reciprocal	−0.128 (0.158)	−0.201 (0.166)
Negatively reciprocal if treated unfairly	−0.015 (0.295)	−0.042 (0.397)
Negatively reciprocal if others treated unfairly	−0.134 (0.320)	−0.552 (0.392)
Individuals	266[a]	158[a]

Notes: Each coefficient is from a separate regression where each of the variables listed in the first column is regressed on a dummy variable which is 0 for the data collected at T1 and 1 for the data collected at T2. The second column reports this dummy's coefficient when using the pooled full samples (all participants assigned to first mover roles in TG or mTG). The third column reports this dummy's coefficient when using the pooled estimation samples. At T2, the estimation sample coincides with the full sample. At T1, I kept in the estimation sample those first movers with minor or no understanding mistakes.

A positive and significant coefficient shows that the respective characteristic is more likely in the sample at T2 compared to the sample at T1. The figures in parentheses are standard errors robust to clustering at the session level. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving. Variables 'Others can be trusted', positive reciprocity and negative reciprocity when others are treated unfairly or when oneself is treated unfairly are measured on a 0–10 scale, where 0 is full disagreement with the statement and 10 is full agreement with the statement.

[a] There are 265 (157) respondents for the regression of gender, as one respondent did not specify their gender.

Table C.3: Minimum acceptable probabilities at T2

	TG	
	Correct answers	Incorrect answers
	58.19 (27.28)	56.90 (25.14)
p-value	0.667	
Observations	62	30
Individuals	31	15
	mTG	
	Correct answers	Incorrect answers
	62.97 (23.74)	60.51 (21.97)
p-value	0.067	
Observations	38	56
Individuals	19	28

Notes: The table shows the average MAP per treatment at T2 for those with correct answers (minor or no mistakes) versus those with incorrect answers to the comprehension questions. Each participant made two decisions. *P*-values are from ranksum tests between the two columns. Standard deviations in parentheses.

Appendix D Checking for order effects

Table D.1: Linear regressions on Minimum Acceptable Probabilities at T2
Interaction between treatment, opponent identity and a dummy for the second decision

	(1)	(2)	(3)	(4)
Baseline: mTG2, ingroup match				
TG2	−6.49	−6.12	−8.81	−8.47
	(6.50)	(6.28)	(6.29)	(6.32)
Outgroup match	2.94	5.73	5.00	7.17
	(6.50)	(6.28)	(6.36)	(6.10)
TG2 × Outgroup match	8.54	6.73	10.71	8.41
	(9.61)	(9.16)	(9.62)	(9.29)
Second decision=1	0.06	2.86	2.13	4.30
	(6.68)	(6.77)	(6.54)	(6.46)
TG2 × Second decision	1.49	−0.31	3.67	1.36
	(10.13)	(9.95)	(9.92)	(9.73)
Outgroup match × Second decision	−3.29	−9.35	−7.42	−12.23
	(12.72)	(12.49)	(12.28)	(11.83)
TG2 × Outgroup match × Second decision	−9.51	−5.43	−13.87	−8.78
	(19.40)	(18.53)	(18.86)	(18.08)
Risk loving (0–10)		−3.20***		−2.98**
		(1.12)		(1.20)
Male		0.13		−1.60
		(5.12)		(5.11)
Constant	60.81***	82.32***	54.58***	71.49***
	(4.28)	(8.60)	(6.86)	(9.82)
Session fixed effects	No	No	Yes	Yes
Adjusted R^2	−0.01	0.05	0.07	0.13
Observations	186	184	186	184
Individuals	93	92	93	92
Sessions	10	10	10	10

Notes: Standard errors clustered at the individual level in parentheses. The sample in models (2) and (4) has one respondent less that that in models (1) and (3), because one participant did not specify their gender. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table D.2: Linear regressions on Minimum Acceptable Probabilities at T2
Interaction between treatment, opponent identity and a dummy for choosing an ingroup member in the hypothetical allocation task: first decisions only

	(1)	(2)	(3)	(4)
Baseline: mTG2, ingroup match				
TG2	−18.16*	−14.31	−20.22**	−15.55
	(10.02)	(10.77)	(9.52)	(10.70)
Outgroup match	−14.56	−10.22	−15.01*	−11.37
	(9.05)	(7.88)	(7.88)	(7.09)
TG2 × Outgroup match	23.62	20.38	26.11*	22.05
	(15.58)	(15.50)	(15.13)	(15.35)
Ticket to ingroup	−9.30	−6.98	−10.19	−6.94
	(7.66)	(7.38)	(7.51)	(7.46)
TG2 × Ticket to ingroup	17.47	12.49	17.58	10.88
	(12.98)	(13.95)	(13.77)	(14.89)
Outgroup match × Ticket to ingroup	29.65**	27.26**	33.03***	30.57***
	(11.97)	(11.36)	(11.83)	(11.44)
TG2 × Outgroup match × Ticket to ingroup	−23.77	−22.56	−25.84	−23.10
	(19.06)	(19.12)	(19.27)	(19.40)
Risk loving (0-10)		−2.73**		−2.40*
		(1.18)		(1.21)
Male		−0.67		−1.28
		(5.16)		(5.64)
Constant	67.29***	81.40***	54.93***	64.15***
	(5.09)	(6.99)	(6.80)	(7.48)
Session fixed effects	No	No	Yes	Yes
Adjusted R^2	0.05	0.09	0.09	0.12
Observations	93	92	93	92
Individuals	93	92	93	92
Sessions	10	10	10	10

Notes: The sample in models (2) and (4) has one respondent less that that in models (1) and (3), because one participant did not specify their gender. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix E Sensitivity analysis: T1 and T2

I use Monte Carlo simulations to estimate (i) how much overlap can be expected between the samples at T1 and T2 and (ii) how much the precision of the estimates of interest can be expected to decrease due to this.

First, I check how likely it is that there is no individual present in both estimation samples. I randomly select 65 identifiers (the size of the estimation sample at T1, in models without session fixed effects) from the pool of 642 potential subjects.[32] I then select 93 identifiers (the size of the estimation sample at T2) from the same pool of 642 identifiers. I count how many identifiers the two samples have in common. I repeat the procedure 1,000,000 times. With the simulation seed used, the number of individuals present in both samples ranges from 0 to 24, with a median of 9 individuals. The probability of no overlap is under 1%. This means one cannot simply consider the coefficients from the regressions on the pooled data set as the true coefficients, but one has to estimate their precision given this high probability of overlap.

Next, I estimate how the overlap between samples at T1 and T2 could influence the precision of changes between T1 and T2, for in- and outgroup members, in MAP_{TG}, MAP_{mTG}, betrayal aversion, and discrimination in betrayal aversion. In 10,000 simulations, I randomly draw from 642 random identifiers a set of 65 identifiers (the size of the estimation sample at T1), which I refer to as "the counterfactual T1 estimation samples". From the same 642 random identifies I then draw a set of 93 identifiers (the size of the estimation sample at T2), "the counterfactual T2 estimation samples". I assign these randomly generated counterfactual identifiers to first movers in the two estimation samples. This creates 10,000 counterfactual ways in which there could exist overlap between the samples at T1 and T2.

For each of these 10,000 cases, I regress the MAP on a treatment dummy interacted with an experiment dummy (0 for T1, 1 for T2) and with an opponent type dummy (in- or outgroup), risk attitudes, gender and a dummy for making a decision for an ingroup opponent first. In a separate model, I include session fixed effects.[33] In both models, I cluster standard errors at the counterfactual individual level.

Table E.1 below shows the mean *p*-value for the tests corresponding to hypotheses 7a–10, with their standard errors and 95% confidence intervals for the most complete specification, which includes session fixed effects. The significance

[32] 642 students registered for the first exam session in the study track from which I recruited participants.

[33] In the model with session fixed effects, the size of the counterfactual estimation samples at T1 is 52. This is the number of respondents with minor/no comprehension mistakes at T1 who were not the only ones in their session to fulfill this requirement.

level of the p-values is not affected by the potential overlap in samples. The mean p-values below and those of Wald tests for equality of coefficients in model (4) in Table 6 tell the same story about the change in behavior between T1 and T2.

Table E.1: Simulation: variation in p-values of hypotheses about behavior change

Hypothesis	Mean p-value	95% confidence interval	
H7a	0.904 061 8	0.904 025 9	0.904 097 7
H7b	0.638 860 5	0.638 732 4	0.638 988 7
H8a	0.085 31	0.085 197 6	0.085 422 4
H8b	0.112 374 3	0.112 245 8	0.112 502 9
H9a	0.002 042 5	0.002 034 7	0.002 050 3
H9b	0.029 064 8	0.029 003 9	0.029 125 6
H10	0.296 897 1	0.296 715 5	0.297 078 6

Notes: The table shows the average p-value for Wald tests of equality of coefficients over 10,000 simulations.