An exploration of trust, betrayal, & social identity

AN EXPLORATION OF TRUST, BETRAYAL, & SOCIAL IDENTITY

The author acknowledges financial support from the Department of Microeconomics & Public Economics and the Graduate School of Business and Economics (GSBE) at Maastricht University, as well as the Research Centre for Education and the Labour Market (ROA).

This book was typeset by the author using LaTeX. The cover art is the work of Edward Hopper, *Sun in an Empty Room* (1963).

Published by ROA
P.O. Box 616
6200 MD Maastricht

ISBN: 978-90-5321-620-0
Printed in the Netherlands by Canon

An exploration of trust, betrayal, & social identity

DISSERTATION

to obtain the degree of Doctor at Maastricht University
on the authority of
the Rector Magnificus Prof. dr. Pamela Habibović,
in accordance with the decision of the Board of Deans,
to be defended in public on

Wednesday, 1 November 2023, at 10:00 hours

by

Maria-Eugenia Polipciuc

Supervisor:	Prof. Dr. Frank Cörvers
Co-Supervisors:	Dr. Raymond Montizaan
	Dr. Martin Strobel
Assessment Committee:	Prof. dr. Bart Golsteyn (Chair)
	Dr. Elena Cettolin, Tilburg University
	Prof. dr. Arno Riedl
	Prof. dr. Kirsten Rohde, Erasmus University Rotterdam

Acknowledgments

As I write this section in my office in Vienna—a stone's throw from the Prater amusement park—, I cannot help but remark that my PhD was quite a ride. There were ups and downs—many of them, which extended over a longer time than I would have imagined. It feels good to be now at the point where I can look back, hopefully a little wiser, certainly a little older.

I was very fortunate to have taken this path in September 2014. It took me a couple of years to find the topic that I would finally write my PhD on and to orient myself in this strange place that is academia. I am grateful to my first supervisor, Frank Cörvers, for giving me the opportunity to find out what research I enjoy doing, even when this meant abandoning projects which had seemed promising in the beginning. This freedom to pursue my interests is what made me stay in academia after the PhD. From Raymond Montizaan, my second supervisor, I've learnt—among many other things—data analysis techniques, how to improve my writing, and many random facts about people who travel first class on Dutch trains and buy art at TEFAF. Thank you for opening my eyes to all of this! Martin Strobel, who became my third supervisor along the way, is an epitome of kindness and research integrity. His expertise, his trust in me and his calming presence have been decisive for my path. I am forever grateful for it all.

I'd also like to thank the other faculty and support staff who have helped with advice, information, or simply with a smile on a rainy day. This applies to people in ROA, the old AE1 and AE2 departments—Esther Soudant, Joyce Gruijthuijsen, Elias Tsakas, Elke Lucas, Silvana de Sanctis, Fleur Keune, Sylvia Beenen, as well as those in institutions I've visited during my PhD: NHH in Bergen and (unofficially) the Leibniz Institute in Dresden. My colleagues in the PhD—many of whom have become friends—have contributed in various ways, from academic discussions to creating a friendly climate. Both are invaluable features of a workplace. Thank you to Kim van Broekhoven, Marion Collewet, Alexander Dicks, Maria Ferreira, Nickolas Gagnon, Alexandra de Gendre, Inge Hooijen, Marloes de Hoon, Frauke Mayer, Merve Özer, Sanne van Wetten, and

Henrik Zaunbrecher. Thank you to my wonderful colleagues in Vienna, who have made going back to work after (during?) the pandemic something to look forward to—Diya Abraham, Miloš Fišar, Tine Grimm, Gergely Hajdu, Christoph Huber, Jakob Moller, Tina Monelyon, Tina Radler, Gerlinde Rattasits, and last but not least, Ben Greiner.

My friends from back home or those scattered all over: Cati, Cristina A., Cristina D. and Gloria—thank you for always making me feel part of the group, even though I've been away for so long. My Econosofia & friends extended family: Ana, Veronica, Anghel and Loredana, Laura, Andrei, Laurențiu, Andreea, it is always a joy to see you, and each of you inspires me. The friends I've picked along the way: Angie, Fanny, Lin, Ramona—even if we're by now far from each other, you are close to my heart. A big thank you to those who made me feel at home during my early days in the Netherlands—Matthijs, Lydia, Huib, Bart, and Raluca.

My family has always been there for me. "Thanks" will always be too little to say, as it only captures a tiny fraction of the gratitude and love we share—but thank you, even so, to my parents, to my in-laws, to Wanda, Dragoș, Gruia, and Darie. Vă mulțumesc pentru dragostea, încrederea și răbdarea voastră fără margini. The most special "thank you" goes to my husband, Cosma, with whom life is simply the best it could ever be.

Thesis summary

The chapters in this thesis are bound not by methodology, but by stemming from the same curiosity: that about how to improve cooperation and social cohesion in the light of diversity. Chapter 1 uses observational data from schools in the United States in the '90s to study how variations in exposure to racial diversity in school might affect turnout and political preferences of young adults. Higher racial diversity in school has a positive effect on turnout in presidential elections seven years later, and a higher share of blacks predicts a higher probability to identify as a Democrat. The effects do not differ significantly by race. The remaining three chapters use an experimental approach. Chapter 2 investigates whether preferences for strategic risk (relative to random risk) in one-shot, two player games depend on whether the players' interests are aligned or not. We find that this matters only if interests are not aligned. In this situation, the opponent faces a trade-off between private and social interest. As a result, one can evaluate their intentions from their action. Chapter 3 studies how betrayal aversion contributes to the difference in trust towards ingroup members versus towards outgroup members. Results indicate that it does not play an important role. This is true not only in the short run, but also over time, even as some participants trust ingroup members more than outgroup members seven months after the groups had been created. Chapter 4 empirically tests an assumption made in the identification of betrayal aversion. Results show that this assumption—that the underlying distribution of risks does not influence risk aversion—does not hold. Overall, the chapters in this thesis contribute to experimental methodology, to the experimental research of trust, and to political behavioral economics.

Contents

List of Figures

List of Tables

Introduction

Both common sense and a vast literature agree that generally humans prefer to cooperate with members of their own social group (known as "ingroup members") than with others ("outgroup members").[1] Depending on the society, different group divides are more relevant: for instance, they could be ethnic, religious, or racial.

The dynamics of intergroup interactions depend—among other things—on the size and composition of the groups (see the related literature in Chapter 1 for several examples). Often one group is in a privileged position compared to other(s), and this may lead to discrimination and intergroup inequality. As our societies become increasingly diverse due to migration (McAuliffe and Triandafyllidou, 2021), it is ever more important to understand what hinders intergroup cooperation and how it can be fostered.

In this dissertation, I examine two aspects of intergroup cooperation and potential ways to enhance it. The first aspect is the long-term effect of exposure to racial diversity in adolescence on civic engagement of young adults as proxied by voting and political preferences (Chapter 1). This work bridges two strands of literature. The first strand provides early evidence that contemporaneous community diversity is negatively correlated with civic engagement (Costa and Kahn, 2003; Dinesen and Sønderskov, 2015; Algan et al., 2016; Martinez i Coma and Nai, 2017; Bellettini et al., 2020), with more mixed results in more recent studies (reviewed in Cancela and Geys, 2016). The second strand brings evidence that events in childhood and adolescence shape one's preferences and beliefs over the lifetime (for a review of the literature on the effects of early

[1] For instance, if we focus on trust towards ingroup members versus outgroup members in experimental settings, many papers find higher trust towards ingroup 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).

exposure in economics, see Malmendier, 2021). My research shows that experiencing diversity early on has a positive impact on a form of civic engagement, voting, during adulthood.

The second aspect focuses on betrayal aversion, which has been identified as a determinant of trusting behavior (Bohnet and Zeckhauser, 2004; Aimone et al., 2015; Fairley et al., 2016; Quercia, 2016; Bacine and Eckel, 2018; Butler and Miller, 2018). Betrayal aversion is the name that has been given to the strategic risk premium which several experimental papers find in trust games. In these papers, participants ask for a higher guarantee that they will receive a high payoff to be willing to trust someone than to take an equally risky bet. In three chapters, I use laboratory and online experiments to study several facets of betrayal aversion. In Chapter 2, I study whether a setting in which there is no scope for betrayal has the potential to make players more willing to face risk generated by another person than to take a risky bet. Should this be the case, as some literature suggests (Bolton et al., 2016; Butler and Miller, 2018), an institution could potentially increase the number of trusting interactions by removing the scope for betrayal in a setting where it exists *a priori*. In Chapter 3, I examine the relationship between betrayal aversion and discrimination in trust, defined as trusting ingroup members more than outgroup members. Chapter 4 analyzes a potentially confounding factor—different beliefs about trustworthiness in the trust game and in the control game—in the standard way in which betrayal aversion has been measured. The work in these chapters suggests that researchers studying betrayal aversion should account for such beliefs. Once this is done, betrayal aversion seems to play less of a role in explaining trusting behavior than previously thought—but more research is needed to confirm these findings. These chapters contribute to improving the measurement of betrayal aversion and to understanding its link to discrimination in trust. As such, they add to the body of knowledge about the underpinnings of trust, which is a prerequisite for many economic and social interactions (Arrow, 1974; Schwerter and Zimmermann, 2020).

I. Outline

Chapter 1

In the 1990s, sociologist Robert Putnam wrote about how the social fabric in

the United States was disintegrating at the same time as communities were becoming more racially, ethnically and religiously diverse (Putnam, 2007). Was diversity to blame for this? In the decades that followed, there was plenty of research in the social sciences on this topic. One (correlational) finding was that in places which were more racially or ethnically diverse, various forms of civic engagement—among which turnout in elections—were lower (Costa and Kahn, 2003; Dinesen and Sønderskov, 2015; Algan et al., 2016; Martinez i Coma and Nai, 2017; Bellettini et al., 2020). Possible explanations advanced in the literature were that members of diverse societies lose interest in the public (and political) sphere. This could be due to fewer meaningful interactions with fellow citizens (Algan et al., 2016), or to fewer benefits from civic behavior for members of one's own group—the smaller the group (at least in relative terms), the lower the incentive to engage in behavior with positive externalities for the community (e.g. Vigdor, 2004).

In this chapter, co-authored with Frank Cörvers and Raymond Montizaan, we ask: **what is the relation between the level of racial diversity when growing up and turnout later in life (a proxy for civic engagement)?** There is a vast literature showing that events in childhood and adolescence are especially important for people's trajectories later on (Malmendier, 2021). Could early exposure to racial diversity overturn the negative correlation between community diversity in adulthood and turnout?

To answer this, we study a representative sample of pupils aged 12–18 from the United States. We calculate an index of racial diversity of their school cohort in the 1994–1995 school year. We examine how small and arguably random changes in racial diversity (as measured by a Herfindahl-Hirschman index) from one cohort to the next are linked to respondents' self-declared turnout and political sympathies seven years later.

Respondents from more diverse cohorts are more likely to be registered to vote and to say they have voted in the most recent presidential elections. The cohort's racial diversity does not influence their political leaning, but having had a higher share of blacks in the cohort increases the chance they identify as Democrats.

We lack data on attitudes and beliefs which would help us point to a precise mechanism. However, we find that interracial friendships are more likely in

more racially diverse cohorts. The positive effects on turnout are also stronger for those who live in more racially diverse neighborhoods as adults. These are indications that with richer data it would be interesting to examine the influence of friendships more closely.

Chapter 2

As mentioned earlier, betrayal aversion is an extra guarantee required to trust someone relative to taking a bet that is equally risky. A possible explanation is that is due to the strategic nature of the risk one faces when trusting: betrayal aversion could "insulate" against the additional discomfort caused by the source of risk being a person rather than nature.

Several studies indicate that in one-shot, two-player games, preferences regarding strategic risk (relative to random risk, such as when taking a risky bet) depend on whether players' payoffs are positively or negatively correlated (for instance, Bolton et al., 2016). If there is an outcome which maximizes both players' payoffs—that is, if players' interests are aligned—, individuals prefer to play with a person rather than take an equiprobable bet (Bolton et al., 2016 find this under risk, Chao, 2018; Chark and Chew, 2015, and Chuah et al., 2016 find this under uncertainty, which refers to unknown probabilities for the possible outcomes). If, however, different outcomes maximize the two players' payoffs, then they prefer the to take the bet rather than play with the human opponent (as shown by many papers on betrayal aversion: Bohnet and Zeckhauser, 2004; Aimone et al., 2015; Fairley et al., 2016; Quercia, 2016; Bacine and Eckel, 2018; Butler and Miller, 2018). In trust games, the second mover's interests are not aligned with the first mover's, which means trust games fall into the second category. The second mover maximizes her payoff by betraying the first mover. Should she do that, she ends up with what is called an additional "temptation payoff".

In this chapter, co-authored with Martin Strobel, we ask: **could an institution which removes the temptation payoff in the trust game make individuals prefer strategic risk to an equiprobable random risk?** If betrayal aversion is a hurdle to some otherwise socially desirable interactions, such an intervention could create the conditions for them to happen.

We run a laboratory experiment with university students and measure their risk aversion to strategic risk and to random risk using two games. The games

are a trust game and a game which only differs from the trust game by having no temptation payoff for the second mover. We use the two games to vary whether players have common or conflicting interests. An important feature which distinguishes our experiment from most related work is that we ensure first movers' beliefs are the same in the two variants of each game (where risk is either random or strategic).

We find betrayal aversion in the trust game: the dislike of strategic risk relative to random risk. In the game without a temptation payoff, there is no significant difference between the players' willingness to take strategic versus random risks.

We conclude that risk aversion in games of aligned interests is not sensitive to the source of risk (random or strategic) once one ensures beliefs are the same in the two variants of a game. We argue that for the source of risk to influence stated risk aversion, the opponent has to face a trade-off between common interest and self-interest, which allows for her action to reveal her intention. This is the case in the trust game. This implies that betrayal aversion is a type of intention-based social preference (as initially proposed in the paper introducing the term "betrayal aversion", Bohnet and Zeckhauser, 2004).

Our results imply that removing the temptation payoff eliminates the scope for intention-based social preferences to make a difference between settings with strategic and those with random risk. An institution removing this payoff would only lead to a mechanical increase in interactions, one due to there being nothing to gain from betraying. We find no additional effect of strategic risk becoming preferred over its random counterpart.

Chapter 3

In many economic settings, trust in the person one is dealing with is indispensable for transactions to take place. If trust in outgroup members is lower and some groups are heavily underrepresented on one side of the market, this can lead to discrimination against members of these groups. Some examples are: migrants or people with disabilities receiving fewer job offers, or racial or ethnic minorities being denied service on peer-to-peer platforms, such as AirBnB for accommodation, or Uber for transportation (Ge et al., 2016; Edelman et al., 2017).

In this chapter, I run two laboratory experiments with university students to

understand **how betrayal aversion contributes to a previously documented ingroup bias in trust** (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). Additionally, I study **the effects of natural group formation on how trust towards in- versus outgroup members and its building blocks evolve over time**.

In this study, students were randomly assigned to peer groups at the beginning of an academic year. I measure their trust in and betrayal aversion towards ingroup (outgroup) members at the beginning of the academic year and seven months later. I find that students trust both types of interaction partners equally at the beginning of the year, and are equally betrayal averse to both. Towards the end of the academic year, there is weak evidence that some trust ingroup members more. However, at this point I find no evidence of betrayal aversion, to neither of the two types of partners. Survey data collected at the end of the experiment indicates that students who are relatively more altruistic to ingroup members also trust them more.

The main finding of this study is that betrayal aversion plays no role in the different trust in the two types of partners. A puzzling question emerges: how come I find no betrayal aversion in the second experiment, while most experiments on the subject do?

Chapter 4

This chapter, co-authored with Martin Strobel, is a short exploration of the puzzle that arose in Chapter 3: why was there no betrayal aversion in the second experiment, while most papers find betrayal aversion? A subtle difference in experimental design between Chapter 3 and many other papers on betrayal aversion could explain the difference, should participants not decide like rational expected utility maximizers.

In many papers which find betrayal aversion (mentioned in the summary of Chapter 3), participants might have imagined a different underlying distribution of the winning chance when confronted with a trusting decision as opposed to risky bet. As an example, they could think that the chance that 60% of potential interaction partners are trustworthy is different from the chance that the high payoff has a probability of 60% in the risky bet. If they were sensitive to the distribution of risks in these situations, this could have made them

behave differently in the two settings (Li et al., 2020). So it could be not the source of risk *per se* that led to different behavior, but the different associated distributions of the risks in the two situations.

In this chapter, we ask: **does the distribution of risks one faces influence stated risk aversion, as measured in studies on betrayal aversion?** To test this, we use an online experiment where we remove strategic considerations. We ask participants how favorable a lottery has to be (how likely its high payoff has to be) for them to prefer its outcome over a sure payoff. We vary the underlying distribution of the probability of the high (low) payoff of the lottery in three treatments.

We find that when the distribution of lotteries is more favorable, participants set higher requirements to prefer a lottery from that distribution over a safe payoff. It is possible that the reference point for how much risk is acceptable is influenced by the underlying distribution of risk: worse prospects lead to accepting more risk, while the opposite is true for better prospects. This pattern runs counter to what would generate a premium similar to betrayal aversion.

A related literature on valuation of goods using the Becker–DeGroot–Marschak mechanism finds similar results: people are more willing to pay a higher price for a good if the distribution of potential prices is more left skewed (for a short review of this literature, see Tymula et al., 2016). In both cases, the more likely a good outcome is (e.g. a low price for the good/a high chance of a good lottery), the less one is willing to accept a bad outcome (a high price/a low chance of a good lottery). Since betrayal aversion is measured using an adapted Becker–DeGroot–Marschak mechanism, it is possible that these effects are relevant in our context as well. Given our findings, we recommend controlling for beliefs about trustworthiness when studying betrayal aversion.

II. Scientific and policy relevance

This dissertation contributes to discussions about (i) the societal impact of educational policies, particularly those regarding racial diversity and about (ii) betrayal aversion as a determinant of trust and the potential for its reduction to enhance economic interactions such as transactions.

The first point is relevant to both academics and policy makers as it broadens the area of effects which could be considered when evaluating educational poli-

cies. Except for the widely considered effects on labor market outcomes, the study in Chapter 1 indicates that policies that impact racial diversity in schools might have an effect on pupils' civic engagement and thus, on social cohesion in the communities in which they live as adults. The findings in this chapter add to the causal evidence underscoring the importance of early socialization for long-term outcomes. Since we examine early socialization in schools, this chapter belongs to a paradigm in behavioral studies which is increasing in popularity: that which focuses on creating evidence to support system-level change rather than individual-level change to address societal issues (Chater and Loewenstein, 2022). According to this paradigm, more studies should examine system-level change, as this type of change is both understudied (relative to individual-level change in behavior) and essential for addressing social challenges such as inequality or discrimination.

The second point is more relevant for academics, as more research is needed to make policy recommendations targeting betrayal aversion. As explained in chapters 2, 3 and 4, many of the previous studies which find betrayal aversion use an identification method that hinges on a questionable assumption. In chapters 2 and 3, we use a cleaner control treatment. In Chapter 2, we study the effects of removing the scope for betrayal by eliminating the temptation payoff that could be gained by betraying. This bolsters interactions by making fewer players betray: a mechanical effect due to a change in incentives. Based on previous literature, we were expecting to see an additional increase in interactions due to individuals "lowering the bar" for being willing to interact with others. However, we do not find such an additional effect. This result provides evidence that betrayal aversion is most likely an intention-based social preference. It also highlights that intention-based preferences require settings in which there is tension between the private and the social interest in order to manifest.

An example of a setting where institutions might want to increase the number of trusting interactions is between members of different social groups, when at baseline one group is systematically discriminated against. For example, imagine members of an ethnic group own the majority of rentals in a community. If they are more likely to rent to members of their own ethnic group, a social planner might consider interventions to increase the chances that members

of other ethnic groups are also offered a rental. Results in Chapter 3 suggest that when it comes to interactions involving trust (such as the rental example above), trusting an ingroup member more than an outgroup member cannot be attributed to differences in betrayal aversion towards the two types of interaction partners. This indicates that it would be more fruitful for the social planner to tackle discrimination in trust by addressing beliefs about trustworthiness and outcome-based social preferences (such as altruism towards in- versus outgroup members). However, this result needs to be replicated with other social groups before making a policy recommendation. Going back to Chapter 1, evidence from the motivating literature suggests that intergroup interactions at an early age are a promising avenue to reduce discrimination.

Chapter 4 also provides evidence useful to fundamental research about trust. Here, we isolate the impact of using a control treatment resting on a strong assumption on how betrayal aversion is identified (this control treatment has been used in many studies on betrayal aversion). The results in this chapter indicate that changes in this treatment have a large impact, but in the opposite direction than we expected. This implies that when studying trust one should always ask participants about their beliefs in the partner's trustworthiness. Failing to do so opens the door to attributing effects to the wrong reasons.

Chapter 1

Peers' race in adolescence and voting behavior

Abstract

Using a representative longitudinal survey of U.S. teenagers, we investigate how peer racial composition in high school affects individual turnout of young adults. We exploit across-cohort, within-school differences in peer racial composition. One within-school standard deviation increase in the racial diversity index leads to a 2.3 percent increase in the probability of being registered to vote seven years later and to a 2.6 percent higher probability of voting six years later. These effects are likely due to positive interracial contact when socialization has long-lasting effects: higher racial diversity in school is linked to more interracial friendships in school and later on.

This chapter is co-authored with Frank Cörvers and Raymond Montizaan.

I. Introduction

Economists have become increasingly aware of the importance of well-functioning institutions, such as the judicial and political systems, in determining economic performance (Costa and Kahn, 2003). Several studies have established that social capital is a key determinant of good institutions and of economic growth (Knack and Keefer, 1997; Alesina and La Ferrara, 2000; Costa and Kahn, 2003; Guiso et al., 2004). A large literature finds that community heterogeneity leads to lower levels of social capital and to fewer interactions among community members, which is potentially problematic in an increasingly diverse world. Community diversity—be it ethnic, racial, or religious—negatively affects participation in social activities (Alesina and La Ferrara, 2000), trust (Costa and Kahn, 2003; Dinesen and Sønderskov, 2015), the quality and quantity of publicly provided goods (Alesina et al., 1999; Vigdor, 2004; Putnam, 2007), the willingness to redistribute income (Luttmer, 2001; Dahlberg et al., 2012), donations (Andreoni et al., 2016), and economic growth (Easterly and Levine, 1997; Montalvo and Reynal-Querol, 2005).

On a more positive note, recent empirical evidence shows that beliefs about members of other racial or ethnic groups evolve with contact and integration (Boisjoly et al., 2006; Finseraas et al., 2019; Rao, 2019; Schindler and Westcott, 2020; Lowe, 2021; Steinmayr, 2021; Corno et al., 2022) and that intergroup trust can be built even after ethnic conflict (Mousa, 2020). Exposure to diversity at an early age can create a common culture that leads to less redistributional conflict among social groups later in life and may reduce transaction costs between individuals from different social groups, by diminishing the social distance between them or by changing their beliefs about members of other groups (Gradstein and Justman, 2000, 2002). Moreover, having more racially diverse friend groups can increase support for affirmative action (Boisjoly et al., 2006).

The current literature which reports mostly negative effects of diversity generally focuses on contemporaneous measures of racial diversity and civic engagement. For this reason, it cannot establish whether there exist positive or negative long-term effects of diversity on civic engagement, as a result of evolving beliefs or intergroup trust due to earlier contact and integration. It therefore

remains an empirical question whether exposure to racial diversity early on has a negative impact on the civic engagement of individuals later in life, even after they have moved to other communities with different racial compositions.[1]

Moreover, the negative effects of diversity on civic engagement are found in rather large communities, making it hard to observe evolving beliefs and intergroup contacts. Examples include counties, cities, or at the very minimum, census blocks (Dinesen and Sønderskov, 2015; Algan et al., 2016; Cancela and Geys, 2016). These findings might differ in smaller units of observation, such as the school cohort level we study, because it is likely that intergroup contact works differently at this lower level of aggregation (Algan et al., 2016).

In this paper, we focus on voting, which is a type of civic engagement whose relationship to community racial composition has been investigated in a number of studies (Costa and Kahn, 2003; Oberholzer-Gee and Waldfogel, 2005; Cancela and Geys, 2016; Shertzer, 2016; Martinez i Coma and Nai, 2017; Bellettini et al., 2020).[2] We address the gaps in the literature by examining the causal long-run impact of the racial composition of one's peers in high school, which is arguably a community in which intergroup contacts are likely to evolve, on voting behavior later in life, and explore channels through which peer racial diversity could play a role. For this, we use the National Longitudinal Study of Adolescent to Adult Health (Add Health). More specifically, we focus on individual voter participation and political orientation. Using the terms employed by (Manski, 1993), this paper aims to identify exogenous effects of peer racial composition in high school on turnout later in life: that is, how an individual's probability to vote varies with exogenous characteristics of her peer group (here, race). Exogenous effects are contrasted with endogenous effects (how an individual's probability to act varies with the group's behavior) and correlated effects (an individual's behavior is similar to that of her group because of their common environment). This method has been validated and used by several studies that have shown that variations in peers' race, gender, ability, language spoken at home, and exposure to family violence influence in-

[1] Contemporaneous measurement of diversity and civic engagement also increases the likelihood that differences in racial diversity and civic engagement between communities are driven by other factors occurring at the time of measurement.

[2] Voting is considered a measure of civic engagement because the individual bears the private costs (time spent voting, time spent informing oneself about the elections, etc.) of an action that has public benefits, at least for their group.

dividual test scores and post-secondary outcomes.[3] The Add Health data set is particularly suited for measuring the long-run effects of racial diversity in high school on voting behavior since it follows multiple student cohorts from the same school into adulthood.

To our knowledge, this is the first study to use this across-cohort, within-school strategy to examine the long-term effects of the racial composition of one's peers in high school on individual voter participation and political orientation later in life. Previous studies on racial diversity focus on the short-run effects and find a negative relationship with turnout (Costa and Kahn, 2003; Martinez i Coma and Nai, 2017; Bellettini et al., 2020).[4] This paper is also closely related to an emerging literature on the effects of school desegregation policies on voting registration, turnout, and political partisanship. The studies of Kaplan et al. (2019), Bergman (2020), and Billings et al. (2021) find mixed effects of the examined policies on one or more of these outcomes. Bergman (2020) analyzes the short- and long-run risks and benefits of a randomized racial desegregation program in elementary schools for minority students. He finds that being offered a transfer to a low-minority share, higher-resource school increased the likelihood to vote, although exclusively for males. Billings et al. (2021), who examine the end of race-based busing in Charlotte-Mecklenburg, North Carolina schools in 2002–2003, and Kaplan et al. (2019), who focus on a policy that bused students in Jefferson County, Kentucky, between 1975–1985, on the other hand, find no effects of these policies on voting registration and turnout. However, both studies report that these policies increased the likelihood to be registered as a Democrat.[5] Both Kaplan et al. (2019) and Billings et al. (2021) use data from Southern states, and the latter study also focuses on a policy implemented decades ago. This raises the question to what extent their results

[3]For example, Hoxby (2000a,b); Angrist and Lang (2004); Gould et al. (2009); Hanushek et al. (2009); Carrell and Hoekstra (2010); Bifulco et al. (2011); Friesen and Krauth (2011); Lavy and Schlosser (2011); Lavy et al. (2012); Black et al. (2013); Merlino et al. (2019); Brenøe and Zölitz (2020); Briole (2021).

[4]A recent review study of aggregate-level research on turnout levels by Cancela and Geys (2016), however, shows more mixed effects of ethnic diversity.

[5]Billings et al. (2021) find that a 10-percentage point increase in the share of minorities in a student's assigned school decreased their likelihood of registering as a Republican 15 years later by 8.8 percent, with the effect being driven by white students. Kaplan et al. (2019) find that white males who had been assigned to be bused in Jefferson County, Kentucky, between 1975–1985 were significantly more likely to be registered as Democrats forty years later.

are generalizable to other states and to exposure to racial diversity in more recent years. A major benefit of our study is that, unlike data from the natural experiments from desegregation policies, which are local, it uses data which are sampled to be representative of the U.S. middle and high school population in 1994–1995. This increases the external validity of our results. Our study also differs in that we measure racial diversity using a Herfindahl-Hirschman index of racial diversity, while the studies on desegregation policies focus on minority shares or on randomized access to schools with a lower share of one's own race. This racial diversity index is the most common way to operationalize diversity in the literature studying the effects of diversity on proxies of civic engagement. As a result, using this index allows us to compare our results with those of previous studies in this literature.

Studying the impact of peers' racial composition in high school on voting behavior later in life is of particular relevance, since the respondents are adolescents when exposed to racial diversity. Their preferences, beliefs and personality traits are still highly malleable (Borghans et al., 2008), so exposure to members of other racial or ethnic groups has the potential to change them considerably. Adolescence is the phase in which fairness and efficiency considerations, which have been proven to be good predictors of political affiliation and decisions (Fisman et al., 2017; Kerschbamer and Müller, 2020), seem to crystallize (Almås et al., 2010). Given the greater malleability of personality traits and preferences during adolescence, schools play a crucial role not only in transmitting knowledge, but also as a socializing force fostering civic engagement (Gradstein and Justman, 2000, 2002). Many studies find that, unlike socialization during adulthood, early socialization has a durable effect on political attitudes and voting behavior (Jennings and Markus, 1977, 1984; Malmendier and Nagel, 2011; Madestam and Yanagizawa-Drott, 2012; Giuliano and Spilimbergo, 2013; Kim and Lee, 2014; Algan et al., 2019; Akbulut-Yuksel et al., 2020). For example, using historical data from post-WWII Germany, Akbulut-Yuksel et al. (2020) show that the expulsion of Jewish professionals had long-lasting detrimental effects on the political interest and participation of Germans who were in their impressionable years (ages six to 23) during the Nazi regime, but not on adults. The authors further demonstrate that these adverse effects can be explained by the social changes brought about by the ex-

pulsions, which led to lower adult socioeconomic status (SES) and lower civic skills for individuals in their impressionable years during that time. Madestam and Yanagizawa-Drott (2012) find similar effects of attending Fourth of July celebrations on political engagement: while the impact of attendance on voter participation and political orientation was long-lasting for young individuals (ages four to 18), there were no significant long-term effects on the political behavior of adults. Hence, should there be long-run positive effects of racial diversity on voting behavior due to changes in beliefs about members of other racial or ethnic groups from contact and integration, we can expect these effects to be strongest for individuals who are relatively young when exposed to racial diversity.

Our most important finding is that exposure to a more racially diverse cohort in high school increases the probability of voting as a young adult. Greater diversity (as measured by a Herfindahl-Hirschman index of racial diversity) results in higher probabilities of being registered to vote seven years later and of having voted in the presidential elections in 2000, six years later. We measure racial diversity at the cohort level within schools; an increase in the cohort racial diversity index by one within-school standard deviation leads to an increase of 1.7 percentage points in the probability of being registered to vote seven years later (2.3 percent of the mean) and to an increase of 1.1 percentage points in the probability of having voted in the presidential elections six years later (2.6 percent of the mean). For comparison, door-to-door canvassing leads to a 7.1 percentage points increase given a turnout base rate of 50% (de Rooij et al., 2009). The prospect of being asked whether they have voted in a survey increases turnout by 0.2 to 0.3 percentage points (DellaVigna et al., 2016; Rogers et al., 2016). In Bergman (2020), being offered to transfer to a majority white school between 1998–2008 increased the probability to register to vote by 1 percentage point and reduced the probability to vote by 3 percentage points in the presidential elections in 2016, but the effects are not significant. Racial diversity in high school, however, does not affect whether and with which party individuals identify as young adults.[6]

[6]Our results might seem contradictory at first with those of the studies evaluating desegregation policies. However, much of the differences in results reflects the choice of the measure of interest. While our paper uses a racial diversity index, the studies of Kaplan et al. (2019) and Billings et al. (2021) analyze the share of minorities, or randomized access to a school with a

We further examine several potential channels through which racial diversity in high school could affect voting behavior positively in the long run. Our estimations show that positive intergroup experiences exist due to racial diversity: respondents in diverse cohorts have more interracial friendships both during high school and 14 years later, in Wave 4. We also find marginal evidence that a more racially diverse cohort in school has an impact on personality traits in Wave 4 (higher extraversion and conscientiousness). Both channels are indicative of an early socialization mechanism driving the change in the probability to be registered to vote and turnout. Our results thus indicate that a broader, long-run assessment of the impact of racial diversity on voting behavior is necessary. The long-term benefits due to more racially diverse friend groups and smaller social distance between these groups could outweigh their short-term costs.

This paper is organized as follows. Section II describes the data and presents descriptive statistics. Section III discusses our empirical strategy. Section IV presents our econometric results. The last section concludes.

II. Data and descriptive statistics

II.A. Data

We use data from Add Health, a school-based longitudinal study of a nationally representative sample of adolescents in the United States. The data were collected in several ways from adolescents, their fellow students, school administrators, parents, siblings, friends, and partners. In Wave 1, an in-school survey was conducted to provide data on the school context and friendship networks. Thereafter, five in-home surveys were held in Waves 1 to 5. Register databases with information on respondents' neighborhoods and communities have been merged with the Add Health data set. In our analyses, we use information from the in-school survey, from the in-home surveys in Waves 1, 3 and 4, as well as from the merged register data sets.

lower share of one's own race than the original school. The results of our robustness analyses using racial shares are in line with Kaplan et al. (2019) and Billings et al. (2021). That is, the share of blacks in one's cohort has no impact on registration or voting—but it does have a positive effect on the probability to identify as a Democrat. Our paper thus complements and is consistent with the existing findings in this emerging literature.

The in-school survey in Wave 1 took place in 1994–1995. Around 90,000 individuals in grades 7 through 12 participated. The in-school survey gathered basic information, such as respondents' gender, race, and parents' level of education. This data enable us to construct the main explanatory variable: the racial diversity index of one's cohort.[7,8]

Add Health uses a clustered sampling design, in which the schools were sampled first and then the pupils in these schools were selected to form the core in-home sample (approximately 17 females and 17 males were selected randomly from each grade for most schools; in 16 of the 132 schools from which students were followed longitudinally, all the enrolled students were selected). This core sample was enhanced with a variety of oversamples based on race, twin status, disability status, and other categories. The enhanced sample (approximately 20,000 individuals in Wave 1) was subsequently followed longitudinally through in-depth interviews at home.

Our dependent variables reflect political attitudes and behavior, which were assessed in the in-home survey in Wave 3. Wave 3 took place in 2001–2002 and tracked approximately 15,000 of the Wave 1 in-home respondents. The great majority of the respondents in Wave 3 were aged 18 to 26. We also use a couple of variables from Wave 4 for robustness checks and for assessing potential mechanisms.

The estimation sample only includes schools with respondents in more than one cohort in both Waves 1 and 3 (a requirement of the estimation strategy; see Section III.A) and non-missing values for a set of control variables.[9] We focus on cohorts with at least 10 respondents in the in-school survey in Wave 1. This leads to a number of 122 schools (423 cohorts) in the estimation sample from the total of 132 schools (551 cohorts) in which students were followed longitudinally. The size of the median cohort among cohorts from which students were followed longitudinally is 154 (minimum 1, maximum 691). The

[7]Students could declare being a member of more than one race. Since for the analysis we had to assign them to one race only, we followed Bifulco et al. (2011) and gave precedence to the answers in the following order: black, Hispanic, Asian, white, and other. For instance, if a respondent claimed to be both white and Hispanic, the respondent would be considered Hispanic in our analysis.

[8]In the in-school survey, 11.92% of those who answered the questions about race selected more than one race. Figure 1.1 in Appendix 1.B shows how our results on voting are influenced by various amounts of measurement error in peers' race.

[9]All the control variables and their definitions are listed in Table 1.12 in Appendix 1.A.

size of the median cohort in school at W1 for respondents in the estimation sample is 164 students (with a mechanical minimum of 10 and a maximum of 691). The median number of respondents from one cohort in Wave 1 who were surveyed in Wave 3 is 23 (minimum one, maximum 462). Of these, the median number of respondents in the in-home survey in Wave 3 from a school cohort in the estimation sample is 24 students (minimum seven, maximum 386).[10] The estimation sample thus obtained consists of 12,070 individuals. Appendix Table 1.13 shows how each restriction contributes to the size of the estimation sample.

II.B. Variable definitions

Dependent variables

The four dependent variables in our analyses are dummy variables from the in-home survey in Wave 3. These measure voting behavior and political partisanship, as follows: 1) the first dummy measures whether the respondents were registered to vote in Wave 3, 2) the second measures whether they voted in the 2000 presidential election (Bush versus Gore), 3) the third measures whether the respondents identified with a political party, and 4) the fourth dummy measures whether the respondents identified with the Democratic Party.[11] It is important to note that identification as a Democrat is not conditional on identification with a party, such that a value of one means the respondent identifies with the Democratic Party and zero means the respondent either does not identify with any party at all or identifies with another party.

Racial diversity index

Most studies examining the effects of diversity on economic outcomes operationalize diversity by an index and/or the share of the minority (or minorities)

[10] See Bifulco et al. (2011) for a more detailed explanation of the sampling process.

[11] In the replication files, we also consider a fifth dependent variable: a dummy for identifying with the Republican Party. We did not include it in the main results discussed in the paper since our measures of racial diversity do not affect it. We mention it briefly in Section V.B, to support the interpretation that there are no effects of SES inequality on political polarization.

of interest.[12] In our main analysis, we use a racial diversity index.[13] To calculate the index, we construct the racial shares for each cohort in a school from the race students declare in the in-school questionnaire of Wave 1. The racial diversity index is computed as one minus the sum of the squared shares of each race possibly present in a cohort (c) within a school (s).[14] We consider five mutually exclusive categories (i) for race, ordered here by group size: white, black, Hispanic, Asian, and other:

$$RacialDiversity_{cs} = 1 - \sum_{i} share_{ics}^{2}$$

The racial diversity index reflects the probability that two individuals chosen at random from a cohort belong to two different racial groups. This index is the most widely used in the literature on diversity (Alesina and La Ferrara, 2005). Since we consider five possible races, this index ranges between zero (only one race is present in the school-cohort combination) and 0.8 (all five races are present in equal proportions). The index can increase in two ways: as the racial shares become more equal and as the number of racial groups in the cohort increases.

II.C. Descriptive statistics

Table 1.1 and Appendix Table 1.14 describe the characteristics of the estimation sample in column (1). Columns (2) and (3) present the means and standard deviations for the white and minority (or nonwhite) subsamples, respectively.

It is immediately apparent that the minority respondents are in high school cohorts with higher proportions of black students and greater racial diversity. They also come from lower-income families and are more likely to live in urban areas, to be slightly older in Wave 1, and to have more pupils in their class. In Wave 3 (seven years after Wave 1), they earn less than their white peers on average and have a slightly lower education.

[12] For example, Alesina and La Ferrara (2000); Costa and Kahn (2003); Bifulco et al. (2011); Algan et al. (2016); Merlino et al. (2019).

[13] Robustness checks in Tables 1.17 and 1.18 in Appendix 1.B use alternative measures of diversity.

[14] This index is known in the literature as the fractionalization index. The index is equal to one minus the Herfindahl-Hirschman index (or one minus the concentration index).

Table 1.1: Summary statistics: main variables

	All		White		Minority	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Main variables						
Is registered to vote	0.738	0.440	0.742	0.437	0.730	0.444
Voted in 2000	0.442	0.497	0.441	0.497	0.444	0.497
Identifies with a party	0.339	0.473	0.334	0.472	0.350	0.477
Democrat	0.183	0.387	0.136	0.343	0.289	0.453
Republican	0.156	0.363	0.198	0.399	0.061	0.239
Racial diversity index	0.373	0.197	0.328	0.180	0.473	0.198
Racial SES Gini	0.362	0.181	0.366	0.183	0.352	0.176
Shares in cohort						
Share white	0.634	0.296	0.763	0.197	0.346	0.276
Share black	0.175	0.242	0.096	0.150	0.352	0.308
Share Hispanic	0.112	0.161	0.073	0.091	0.199	0.233
Share Asian	0.039	0.074	0.028	0.051	0.064	0.106
Share other	0.039	0.034	0.039	0.028	0.039	0.045
Own race						
White	0.691	0.462				
Black	0.161	0.367				
Hispanic	0.100	0.300				
Asian	0.035	0.184				
Other	0.013	0.114				
N	12,070		6,692		5,378	

Notes: Summary statistics are calculated using Wave 3 longitudinal weights, which aim to produce a representative sample of individuals who were surveyed in both Waves 1 and 3.

In terms of voting behavior, minorities are also slightly less likely to be to be registered to vote, as well as to have voted in the 2000 presidential elections. In the full estimation sample, approximately 44 percent declared having voted in the presidential elections in 2000 and about 74 percent were registered to vote when the data collection of Wave 3 took place (August 2001 to April 2002). We do observe a substantial difference between whites and minorities with respect to party identification. Minorities are twice as likely as whites to identify with the Democratic Party (28.9 percent versus 13.6 percent).

Table 1.1 also presents the within-cohort racial SES Gini index, which has similar means for white and minority respondents.[15]

Another variable showed in Appendix Table 1.14 is a cohort dummy indicating whether the pupils are grouped by ability in English. Roughly half of the respondents are in cohorts grouped by ability, with minorities being slightly more likely to be in such cohorts.[16]

III. Identification and estimation strategy

III.A. Empirical implementation

To examine the causal long-run impact of the racial composition of peers in high school on voting behavior and political partisanship, we use an analytical strategy that eliminates the bias created by families or students systematically

[15]The racial SES Gini index is a Gini coefficient of the inequality in mothers' education by race in the cohort. We use the formula of Alesina et al. (2016), where, for each race, we consider two educational categories of mothers: with and without a college degree. The Gini coefficient for a cohort with *n* racial groups with a share y_i of college-educated mothers in group *i*, where *i* to *n* are indexed in nondecreasing order ($y_i \leq y_{i+1}$), is computed as

$$Gini = \frac{1}{n}[n + 1 - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}]$$

The coefficient is equal to zero if there is only one racial group or if the shares of college-educated mothers are equal for all the races represented in the school cohort. In a cohort with m races, the coefficient is highest when the share of college-educated mothers is zero for $m-1$ races and one for the remaining race. In this case, the Gini coefficient is $(m-1)/m$. In this paper, we consider five races, so the Gini coefficient has an upper bound of 0.8.

[16]The information on whether a certain cohort in a school is grouped by ability is provided in Wave 1 by the school principal. For each cohort in the school, the principal answered either yes or no to the following question: "For English or language arts, does your school group classes according to ability or achievement?". In the United States, language arts refers to the area of the curriculum in which students are taught the range of skills needed to become proficient in using a language (Moreau, 2011).

sorting into schools, which can be observed in Table 1.1. We exploit idiosyncratic variation in cohort composition between adjacent grades within schools. We thus assume that, conditional on attending a certain school, the cohort composition a pupil faces is as good as random. This method was pioneered by Hoxby (2000b) and is widely used in education economics in pre-university peer studies, where random assignment is rarely feasible.[17]

To implement this strategy, we estimate a reduced-form equation using a linear probability model in which the outcome of an individual student is a linear function of the student's own observable characteristics, the mean characteristics of all the students in the same cohort and school, cohort fixed effects, school fixed effects, and school-specific linear trends. We include the racial SES Gini index and the ability grouping dummy among the controls in our analyses. We thus account for the possibility that the differences in voting behavior are not caused by racial diversity itself, but rather by exposure to the socioeconomic inequality associated with the racial diversity within school cohorts. Should the coefficient of the racial diversity index be significant, this would indicate that other facets of racial diversity than disparities in mothers' education by race matter for the outcome of interest.[18,19] As for ability grouping, if this practice clusters students by race, the chances of interracial contact would be reduced in grouped cohorts. Since positive intergroup contact is a potential mechanism through which racial diversity positively affects voting behavior, not controlling for ability grouping could lead to an underestimation of the impact of racial diversity on voting behavior.

The reduced-form equation looks as follows:

$$\begin{aligned} PolOutcome_{ics} = \alpha + \beta_0 RacialDiversity_{cs} + \beta_1 X_{cs} + \beta_2 X_i + \\ + \delta_c + \phi_s + \lambda_s C + \epsilon_{ics} \end{aligned} \tag{1.1}$$

for individual i, cohort c, and school s. The variable $PolOutcome_{ics}$ is one of

[17] See, for instance, Carrell and Hoekstra (2010); Bifulco et al. (2011); Lavy and Schlosser (2011); Lavy et al. (2012); Black et al. (2013).

[18] The correlation between the racial diversity index and the racial SES Gini index in the 423 cohorts in the estimation sample is –0.012.

[19] We have also used alternative racial inequality Gini indices as control variables in additional robustness analyses, such as a racial SES Gini index for fathers' education, one for mothers' employment status, and one for fathers' employment status. None of these indices changes the magnitude of the coefficient of the racial diversity index for voting or being registered to vote significantly. Results are available on request.

five dummy variables from Wave 3, indicating individuals who, respectively, are currently registered to vote, voted in the presidential elections in 2000, identify with a party, and identify with the Democratic or with the Republican Party. Students' race is one of five mutually exclusive categories reported in the in-school questionnaire: white, black, Hispanic, Asian, and other. The variable $RacialDiversity_{cs}$ is the racial diversity index; X_{cs} includes the racial SES Gini index and the indicator for ability grouping; X_i is a set of individual characteristics; δ_c is a cohort fixed effect common to all schools; ϕ_s is the school fixed effect, which ensures that we compare cohorts within a school; C is an indicator variable for the student's cohort in Wave 1, which is allowed to vary by school; and $\lambda_s C$ allows for the possibility of school-specific linear trends in racial composition. As Bifulco et al. (2011) point out, not controlling for trends is problematic if, for instance, parents decide to enroll their children in a school based on the observed trend in the school's racial composition. The term ϵ_{ics} is a random error term. Standard errors are clustered at the school level.[20]

III.B. Identifying variation

Two conditions must be met for the resulting estimates to have a causal interpretation: (i) there should be sufficient variation in cohort composition within schools and (ii) the source of variation should be plausibly random. With regard to the latter issue, it is important to note that many of the mechanisms through which racial composition can influence outcomes are constant across cohorts in the same school. For example, a school's ability to acquire resources and parents' decisions to place their children in a school are likely influenced by the composition of the school as a whole, instead of that of a particular cohort. Our identification strategy relies on within-school variation in cohort composition so our estimates will not capture any effect that the student composition of the school as a whole has on individual outcomes. Below we present tests of conditions (i) and (ii).

Table 1.2 addresses condition (i): the top panel shows the standard deviation of the main explanatory variables in the estimation sample; the middle

[20]We also conducted robustness analyses in which we omitted the racial SES Gini index and the ability grouping indicator. Our results are robust to this exercise; the coefficients of the racial diversity index remain similar for both being registered to vote (p-value< 0.01) and having voted (p-value< 0.1). The results are available on request.

Table 1.2: Variation in cohort composition measures after removing cohort and school fixed effects and trends

	Panel A Raw cohort variables				
	N	Mean	Standard deviation	Min	Max
Share white	12,070	0.634	0.296	0.000	1.000
Share black	12,070	0.175	0.242	0.000	1.000
Racial diversity index	12,070	0.373	0.197	0.000	0.777
Racial SES Gini	12,070	0.362	0.181	0.000	0.752
	Panel B Residuals after removing school and cohort fixed effects				
	N	Mean	Standard deviation	Min	Max
Share white	12,070	0.000	0.034	−0.271	0.214
Share black	12,070	0.000	0.026	−0.295	0.126
Racial diversity index	12,070	0.000	0.040	−0.141	0.358
Racial SES Gini	12,070	0.002	0.121	−0.428	0.414
	Panel C Residuals after removing school and cohort fixed effects and trends				
	N	Mean	Standard deviation	Min	Max
Share white	12,070	0.000	0.025	−0.182	0.176
Share black	12,070	0.000	0.017	−0.170	0.192
Racial diversity index	12,070	0.000	0.031	−0.123	0.253
Racial SES Gini	12,070	0.002	0.086	−0.346	0.384

and bottom panels show the variation of the residuals obtained by regressing the respective explanatory variable on school and cohort dummies (middle panel) plus school linear trends (bottom panel). Most of the variation in racial composition is due to differences between schools. Although the within-school variation represents only a small fraction of the total variation in the racial diversity index (16 percent after removing school and cohort fixed effects and linear school trends), this is sufficient to reasonably estimate the effects of small changes in cohort composition. Previous studies that use these data and method come to a similar conclusion (Bifulco et al., 2011; Merlino et al., 2019). The magnitudes reported in the middle and bottom panels are also in line with those in Bifulco et al. (2011).

To address condition (ii), we first run balancing tests, the results of which are shown in Table 1.3. These tests check whether deviations from school-specific fixed effects or trends in cohort composition as measured by the racial diversity index are associated with deviations in student background characteristics (except for one's own race, which is addressed by separate tests, described in the next paragraph). We regress predetermined Wave 1 background characteristics on the racial diversity index and on dummies for own race and own grade (column (1)); we then add school fixed effects (column (2)) and school linear trends (column (3)). None of the coefficients remain significant after the addition of school fixed effects and school linear trends.[21]

Second, if variation in cohort composition is as good as random, a pupil's race should not be correlated with that of his or her peers. To test whether this is indeed the case, we need to consider all the students in cohorts from which there is a student in our estimation sample. However, one cannot simply regress an individual's own race on peers' race to test this: since an individual is always excluded from their peer group, this mechanically creates a negative correlation between the two variables, even in the presence of random vari-

[21] An anonymous referee suggested that the court-ordered school desegregation measures might have affected the share of black students across cohorts and hence the variation of the racial diversity index. We are not able to check whether the schools in the sample are or were previously subject to such court orders. However, similar to Merlino et al. (2019, p. 672), we note that Lutz (2011) shows that the expiration of the orders is not correlated with other trends, which could have affected our identification strategy. If we look at the estimation sample specifically, Table 1.3 shows that the racial diversity index and the racial shares at the neighborhood level are not significantly correlated. This implies that the across cohort variation we observe does not capture other changes in racial composition at the neighborhood level.

Table 1.3: Balancing tests: Racial diversity index

	OLS (1)	School fixed effects (2)	School fixed effects + trends (3)
Individual characteristics W1			
Male	0.041	0.119	0.163
	(0.039)	(0.122)	(0.161)
Age	−0.130	−0.196	−0.287
	(0.077)	(0.165)	(0.215)
GPA	−0.200	−0.009	−0.195
	(0.109)	(0.207)	(0.190)
Ability grouping	0.316	0.055	−0.227
	(0.195)	(0.212)	(0.392)
Family characteristics W1			
Mother's education (years)	0.258	−0.022	0.736
	(0.385)	(0.524)	(0.626)
Log of family income	0.058	0.013	0.040
	(0.033)	(0.048)	(0.062)
English not spoken at home	0.015	0.000	0.000
	(0.026)	(0.032)	(0.034)
Lives with both biological parents	−0.118	−0.061	−0.034
	(0.049)	(0.114)	(0.162)
Parent civically engaged	0.043	0.156	0.088
	(0.061)	(0.129)	(0.164)
Missing parent information	0.072	0.139	0.155
	(0.055)	(0.150)	(0.207)
Neighborhood characteristics W1			
Share less than high school	−0.102	−0.002	−0.038
	(0.047)	(0.025)	(0.035)
Share with bachelor's degree	0.134	−0.018	0.010
	(0.043)	(0.025)	(0.037)
Share votes for Democratic candidate 1992	0.039	−0.006	−0.004
	(0.040)	(0.005)	(0.007)
Share blacks	0.043	−0.039	−0.028
	(0.051)	(0.035)	(0.043)
Share Hispanics	0.091	−0.010	−0.004
	(0.039)	(0.016)	(0.014)
Share Asians & other	0.113	0.008	−0.004
	(0.022)	(0.008)	(0.008)
Share below poverty level	−0.023	0.003	−0.019
	(0.039)	(0.020)	(0.026)
Urban area	0.936	0.055	−0.094
	(0.159)	(0.064)	(0.097)
Other			
Racial SES Gini	−0.177	0.049	0.173
	(0.067)	(0.239)	(0.275)
N	12,070	12,070	12,070

Notes: Each coefficient is from a separate regression where each of the variables listed (measured in Wave 1) is regressed on the racial diversity index with controls including own race and own grade dummies (1), plus school fixed effects (2) and school linear trends (3). We report the coefficient of the racial diversity index. The figures in parentheses are standard errors robust to clustering at the school level. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. Wave 3 longitudinal weights are used.

ation.[22] Guryan et al. (2009) propose correcting for this bias by additionally controlling for all potential peers, that is, by controlling for the racial composition of the school as a whole. We regress each race dummy on the racial diversity index of peers (excluding oneself) in the cohort and in the school, with grade and school fixed effects and school linear trends. The upper panel in Appendix Table 1.22 presents the results: in each column, the dependent variable is a dummy for being of a certain race. Caeyers and Fafchamps (2020) propose a different approach: regressing a transformed dummy for being of a certain race on the respective racial share of others in the cohort. The race dummy is transformed by subtracting the exclusion bias which creates the artificial negative correlation. The lower panel in Appendix Table 1.22 shows the results for all races considered, also including grade and school fixed effects and school linear trends. From these tables, we conclude that respondents' race is not systematically correlated with that of their peers as measured by the racial diversity index at a significance level of $\alpha = 0.05$.

Third, we conduct Monte Carlo simulations to further test whether the within-school variation observed in the racial diversity index is consistent with a random process. We use two methods to flag those schools from the sample where this is not the case. For the first method, we compute the shares of each race present in the school. Then, for each student in the school, we randomly generate a counterfactual race using a multinomial distribution function with the probabilities of being of a certain race equal to the true racial probabilities at the school level. We repeat the process 1,000 times and compute a confidence interval for the racial diversity index, calculated using the generated counterfactual racial shares in each cohort within each school. We then flag those schools where the true average racial diversity index at the school level does not fall within the 95 percent confidence interval. This method has been used previously by Lavy et al. (2012).

The second method is very similar to the first, but it simulates assignment to a certain cohort in a school using the multinomial distribution of grades within the school. We compute the racial shares in the counterfactual grades and again flag schools where the resulting racial diversity index does not fall within the

[22]This is true of the correlation with the share of peers of the same race. We focus on racial diversity as measured by the racial diversity index, so we adjust the tests accordingly.

95 percent confidence interval. This method has been used by Bifulco et al. (2011).

With our choice of the simulation seed, the first method flags 13.1 percent of the schools in the estimation sample (16 out of 122 schools, comprising 9.2 percent of the students in the main estimation sample), while the second one flags 5.7 percent (seven out of 122 schools, where 4.4 percent of the students in the main sample were enrolled).[23]

A fourth test that we perform to check whether the observed within-school variation in the racial diversity index is consistent with a random process is based on Feld and Zölitz (2017). The authors show that, if the variation one exploits is nonrandom, measurement error in the explanatory variable can create an upward bias in its coefficient. We follow Merlino et al. (2019) and introduce various amounts of measurement error in race. We then examine the resulting pattern of the coefficient of the racial diversity index in two regressions: that of the probability of being registered to vote and that of the probability of having voted in 2000, using the most complete specification in Table 1.4 (presented in the next section).[24] Appendix Figure 1.1 plots the coefficients of the racial diversity index with measurement error from these regressions. Adding error attenuates the coefficients. This supports the assumption that the variation we exploit is as good as random.

A fifth and final test is a permutation test, as discussed in Guryan et al. (2009, p. 48) and Caeyers and Fafchamps (2020, p. 11–12). It works as fol-

[23]As a robustness check, we rerun the main regressions on the restricted samples, which contain only respondents from the unflagged schools. Appendix Tables 1.15 and 1.16 present the results. Compared to results in Table 1.4, the magnitude and significance level of the coefficient of the racial diversity index for being registered to vote are virtually unchanged. For having voted in 2000, the coefficient has the same magnitude, but is insignificant in Table 1.15, and it is higher in magnitude and highly significant in Table 1.16. These results indicate that the variation we exploit is quasi-random.

[24]We introduce measurement error in race in the following way: in the in-school data set, we run a multinomial logit regression of individuals' own race (which can take on five values) on school and grade fixed effects and school linear trends, with standard errors clustered at the school level. We calculate the predicted probabilities of being of each race. We then generate a new variable that takes the value of each race with a probability equal to the predicted probability of being of that race. Then, we generate a new race variable for each level of measurement error considered, denoted me: 0 percent, 5 percent, 10 percent, 25 percent, 50 percent, 75 percent, 90 percent and 100 percent. This variable takes the value of the true race in $100 - me$ of the cases and the value of the new variable in *me* of the cases. We repeat the process 1,000 times for each error level in the measurement of race. We then construct racial diversity indices and racial SES Gini indices for all levels of measurement error in race.

lows: we first regress each of the five race dummies on the share of others in the cohort belonging to the same race and school fixed effects (first specification) plus school linear trends (second specification). We cluster standard errors at the school level. Because of exclusion bias, even if the allocation to cohorts in a school is as good as random after controlling for school fixed effects, with or without school linear trends, the coefficient for the share of others in the cohort is expected to be negative and significant. To check the assumption of random assignment we shuffle cohorts within schools, and thus create counterfactual cohort assignment. We then regress each of the five race dummies on the counterfactual share of others of the same race in the cohort and on school fixed effects (with or without school linear trends). We repeat this process 1,000 times and store the coefficients of the counterfactual cohort composition. For each of the five races, we then check what share of these coefficients is either above the absolute value of the true coefficient or below minus its absolute value. This share is the p-value of the test of random peer assignment for that specific race. For neither of the five races can we reject the null hypothesis of random peer assignment, either with or without including school linear trends.[25]

We conclude that the tests in this section do not reject our hypothesis that, once we control for grade and school fixed effects and school linear trends, residual deviations in the cohort racial diversity index are as good as random.

IV. Results

IV.A. Voting behavior

Table 1.4 presents evidence on the impact of racial diversity in one's high school cohort on voting behavior as a young adult.

[25] In the specification without school linear trends, the number of p-values which are greater in absolute value than the actual p-value for share of white (black/Hispanic/Asian/other race) others in the cohort is 942 (795/990/810/1000). In the specification with school linear trends, the values are 1000 (985/260/976/892).

Table 1.4: Voting behavior in Wave 3

Dependent variable:	Registered to vote			Voted in 2000		
	(1)	(2)	(3)	(4)	(5)	(6)
Racial diversity index	0.552	0.542	0.541	0.379	0.369	0.365
	(0.146)	(0.146)	(0.147)	(0.170)	(0.166)	(0.167)
Racial SES Gini	−0.029	−0.035	−0.036	−0.059	−0.061	−0.060
	(0.076)	(0.074)	(0.074)	(0.052)	(0.050)	(0.051)
Share mothers with college degree	0.144	0.109	0.111	0.263	0.209	0.211
	(0.139)	(0.138)	(0.140)	(0.168)	(0.164)	(0.165)
Ability grouping	0.001	0.005	0.001	0.085	0.087	0.085
	(0.031)	(0.028)	(0.030)	(0.038)	(0.038)	(0.039)
Black	0.051	0.068	0.059	0.058	0.087	0.064
	(0.019)	(0.019)	(0.021)	(0.023)	(0.023)	(0.021)
Hispanic	−0.042	0.001	0.012	−0.089	−0.045	−0.034
	(0.024)	(0.022)	(0.023)	(0.019)	(0.020)	(0.020)
Asian	−0.130	−0.097	−0.091	−0.160	−0.135	−0.134
	(0.037)	(0.036)	(0.036)	(0.035)	(0.029)	(0.029)
Other	−0.032	−0.001	0.000	−0.122	−0.093	−0.097
	(0.047)	(0.046)	(0.048)	(0.048)	(0.049)	(0.051)
Constant	−0.728	−0.828	−0.872	−1.241	−1.370	−1.282
	(0.226)	(0.224)	(0.244)	(0.258)	(0.248)	(0.269)
Individual characteristics	✓	✓	✓	✓	✓	✓
Family characteristics		✓	✓		✓	✓
Neighborhood characteristics			✓			✓
N	12,070	12,070	12,070	12,070	12,070	12,070

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

To interpret the coefficient of the racial diversity index, it is useful to scale it by the within-school standard deviation (0.031, from Panel C in Table 1.2).[26] The dependent variable in columns (1) to (3) is being currently registered to vote. In columns (4) to (6), the dependent variable is having voted in the presidential elections in 2000. In columns (1) and (4), we control for individual characteristics measured in Wave 1. In columns (2) and (5), we additionally control for family characteristics in Wave 1. Columns (3) and (6) show the results when we also control for Wave 1 residential neighborhood characteristics. Our estimates indicate that greater racial diversity in one's cohort in high school as measured by the racial diversity index leads to a higher probability of being registered to vote and a higher probability of having voted in the 2000 presidential elections.

The estimate in the most comprehensive specification in column (3) of Table 1.4 shows that an increase in the racial diversity index by one within-school standard deviation leads to an increase in the probability of being registered to vote of approximately 1.7 percentage points. This represents an increase of 2.3 percent relative to the unconditional probability of being currently registered to vote (73.8 percent, shown in Table 1.1). If we compare the point estimates of the racial diversity index across columns (1) to (3), we further observe that the size of the coefficient is robust to the inclusion of individual, family, and neighborhood characteristics in Wave 1.

Columns (4) to (6) of Table 1.4 show the effects of the racial diversity index on actual voting behavior in the presidential elections in 2000. When using estimates in column (6), we find an increase in the racial diversity index by one within-school standard deviation increases the voting probability by 1.1 percentage points. This represents an increase of 2.6 percent relative to the unconditional probability of voting (36.5 percent). One's own race is also sig-

[26]This would indicate by how much an increase by one standard deviation affects the dependent variable. For the racial diversity index, an increase by one within-school standard deviation (0.031, from Panel C in Table 1.2) from the median racial diversity index (0.393) in a cohort of median size in the estimation sample (164 students) could be achieved in one of the following ways (the list is not exhaustive):

– In a school with only two races present, 120/44 leads to an index of 0.393. A way to increase the index by approximately 0.031 is to make it 114/50.

– In a school with five races present, 125/25/7/6/1 leads to an index of 0.393. Changing it to 122/22/7/7/6 would increase the racial diversity index by approximately one within-school standard deviation.

nificantly related to the probabilities of being registered to vote and of having voted in the 2000 elections. Ceteris paribus, black respondents are more likely to be registered and to have voted than white respondents, while both probabilities are lower for Asians.[27]

Our results therefore suggest a positive causal long-run impact of the racial composition of one's peers in school on voting behavior later in life, even after controlling for own race. The effects we find are sizable compared to previous estimates of the (short-term) determinants of political participation. For example, de Rooij et al. (2009) report that door-to-door canvassing increases turnout by 7.1 percentage points. Furthermore, DellaVigna et al. (2016) and Rogers et al. (2016) find that the turnout rate of people who expect to be asked whether they have voted is 0.3 and 0.2 percentage points, respectively, higher than for those who do not expect to be asked.[28]

IV.B. Political partisanship

In addition to information on voting behavior, Add Health also surveyed respondents about their political attitudes in Wave 3. We have information on whether and with which party people identify. By examining the effects of early age diversity on political partisanship, we can check whether increased

[27]The coefficients in Table 1.4 increase slightly and remain significant at 1% level for being registered to vote, and at 5% for having voted, if we do not include either individual characteristics, family characteristics or neighborhood characteristics. We also check whether the coefficients of the racial diversity index and of the racial SES Gini index are robust when not including the other of the two. This is indeed the case. These results are available in Web Appendix Tables 1 and 3. The Web Appendix is available at: `https://www.dropbox.com/s/r3r3ssufoclyttl/WebAppendix_Peers.pdf?dl=0`

[28]There exists a question about turnout in Wave 4 as well. We decided not to include turnout as measured in Wave 4 as a dependent variable in the main analysis since it is not directly comparable to the turnout variables in Wave 3. First of all, turnout rates in primary/statewide elections (to which the question in Wave 4 refers) are much lower than in presidential elections, e.g. 17.2% among those aged 18—24 and 27.1% among those aged 25—34 years old for the 2002 elections (as reported here: `https://www.census.gov/data/tables/2002/demo/voting-and-registration/p20-552.html`). This is approximately half of the official turnout rate in presidential elections for the same age groups. Second, the question in Wave 4 asks about voting habits in general in these elections rather than about turnout in a specific election. Third, there is attrition between Waves 3 and 4 which means selection might affect results. These reasons make it problematic to compare answers about turnout in the two waves. We present results regarding turnout as measured in Wave 4 in Web Appendix Table 9. The dependent variable is a dummy which is 0 if the respondent said they never vote (32% of all answers) and 1 otherwise. The coefficients of the racial diversity index are close to zero and insignificant.

voting is due to political preferences becoming more alike or, on the contrary, more polarized.[29] It is, however, not clear a priori how these political views are affected by greater racial diversity in the school cohort as measured by a racial diversity index. In both parties, numerous factions cover a wide range of the political spectrum. On some policy issues, there could therefore be overlap between certain factions from the two parties. Furthermore, there are substantial differences within the national and local divisions of these parties. Whether—and how—exposure to racial diversity measured as an index influences political partisanship is thus an empirical question.

[29]Using data from the General Social Survey, Oberholzer-Gee and Waldfogel (2001) document that the political preferences of black and white individuals in the United States differ substantially, with the former being more liberal.

Table 1.5: Political partisanship in Wave 3

Dependent variable:	Identifies with a party			Identifies as a Democrat		
	(1)	(2)	(3)	(4)	(5)	(6)
Racial diversity index	−0.112	−0.124	−0.127	−0.151	−0.155	−0.153
	(0.192)	(0.190)	(0.191)	(0.133)	(0.133)	(0.132)
Racial SES Gini	0.070	0.067	0.066	0.108	0.105	0.104
	(0.053)	(0.051)	(0.052)	(0.035)	(0.034)	(0.034)
Share mothers with college degree	0.245	0.210	0.212	0.094	0.084	0.083
	(0.196)	(0.196)	(0.195)	(0.128)	(0.126)	(0.127)
Ability grouping	−0.040	−0.036	−0.037	0.002	0.005	0.005
	(0.051)	(0.051)	(0.051)	(0.038)	(0.039)	(0.038)
Black	0.095	0.114	0.116	0.222	0.225	0.215
	(0.023)	(0.024)	(0.026)	(0.024)	(0.025)	(0.026)
Hispanic	−0.075	−0.043	−0.034	0.011	0.020	0.027
	(0.026)	(0.026)	(0.025)	(0.024)	(0.021)	(0.021)
Asian	−0.142	−0.125	−0.125	−0.030	−0.029	−0.032
	(0.030)	(0.033)	(0.033)	(0.031)	(0.035)	(0.035)
Other	−0.042	−0.020	−0.019	0.042	0.048	0.045
	(0.059)	(0.058)	(0.058)	(0.042)	(0.041)	(0.041)
Constant	−0.458	−0.554	−0.528	−0.397	−0.377	−0.354
	(0.202)	(0.209)	(0.236)	(0.142)	(0.148)	(0.188)
Individual characteristics	✓	✓	✓	✓	✓	✓
Family characteristics		✓	✓		✓	✓
Neighborhood characteristics			✓			✓
N	12,070	12,070	12,070	12,070	12,070	12,070

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.5 reports the results. In this table, "Identifies as a Democrat" is 1 if the respondent identifies as a Democrat and 0 otherwise (the zeros include those who do not identify with any party). This table is constructed similarly to Table 1.4, by adding the control variables mentioned in the table caption. We find that the racial diversity index in the school cohort is not significantly related to whether people indicate that they identify with a party, nor to whether they identify with the Democratic or Republican Party.[30] We do, however, find that one's own race is significantly related to political partisanship. Black individuals are more likely than white individuals (the baseline category) to identify with a party, and also more likely to identify with the Democratic Party. Asians identify significantly less with a specific party. It is also noteworthy that greater SES racial inequality in high school is linked to a higher probability to identify with the Democratic Party. The effect of the racial SES Gini is higher if we only estimate it for identifying as a Democrat conditional on identifying with a party (0.172, with a standard error of 0.070, compared to 0.104, with a standard error of 0.034 in Table 1.5, in a sample of 4,372 respondents).

IV.C. Underlying mechanisms

In this section, we explore a potential mechanism: that experience of early racial diversity could mitigate the negative effects of racial diversity in adulthood on turnout. We also present tentative evidence of two potential channels through which this could work: interracial friendships and personality changes in adolescence.

Early diversity mitigates the effects of later diversity

It is possible for both a negative correlation between contemporaneous racial diversity and turnout and a positive effect of racial diversity in adolescence on turnout to exist. This could happen if racial diversity early on acts as a mitigating force against the negative effects of racial diversity later in life. If this were true, the positive effect of early racial diversity on turnout should be higher for those living in Wave 3 in neighborhoods that are more diverse.[31]

[30] Results regarding identification with the Republican Party are available on request.

[31] Since where people live in Wave 3 is a choice, such evidence would only be circumstantial, as it could also reflect selection effects (those who are more likely to vote when living among

We first test the assumption that there exists a negative correlation between contemporaneous diversity and turnout. In Table 1.6, we regress the dummy for having voted in 2000 on a racial diversity index computed at the Wave 3 neighborhood level. We compute this index using racial shares at the block group level provided by Add Health. A block group has approximately 1,000 inhabitants in 2000. Add Health defines a block group as follows: it is a "subdivision of a census tract... [and] the smallest geographic unit for which the census bureau tabulates sample data. A block group consists of all the blocks within a census tract with the same beginning number." In column (1), additional covariates include race dummies, gender, age in Wave 1, the share of college-educated mothers in the cohort, the share of males in the cohort, grade and school fixed effects. In column (2), we add school linear trends. Column (3) adds family characteristics in Wave 1 and column (4) adds neighborhood characteristics in Wave 1 (which we list in Appendix Table 1.14). In all four specifications, the coefficient of the neighborhood racial diversity index in Wave 3 is negative and highly significant. This supports our assumption of contemporaneous diversity being negatively correlated with turnout.

diverse others could also be more likely to live in racially diverse neighborhoods).

Table 1.6: Negative correlation between contemporaneous racial diversity and turnout

Dependent variable:	Voted in 2000			
	(1)	(2)	(3)	(4)
Neighborhood RDI in Wave 3	−0.115	−0.120	−0.115	−0.111
	(0.039)	(0.040)	(0.039)	(0.038)
Share mothers with college degree	0.124	0.304	0.256	0.256
	(0.157)	(0.180)	(0.177)	(0.178)
Black	0.066	0.068	0.096	0.070
	(0.024)	(0.024)	(0.024)	(0.023)
Hispanic	−0.079	−0.081	−0.040	−0.029
	(0.020)	(0.019)	(0.020)	(0.020)
Asian	−0.138	−0.147	−0.124	−0.125
	(0.034)	(0.034)	(0.029)	(0.029)
Other	−0.134	−0.131	−0.106	−0.111
	(0.045)	(0.047)	(0.048)	(0.050)
Constant	0.325	−1.070	−1.203	−1.131
	(0.198)	(0.241)	(0.234)	(0.256)
Individual characteristics	✓	✓	✓	✓
Family characteristics			✓	✓
Neighborhood characteristics				✓
N	11,822	11,822	11,822	11,822

Notes: This table reports OLS estimates. Controls include school and grade fixed effects (model (1)) and school linear trends (models (2), (3), and (4)). The omitted category for own race is white. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.7: Early diversity mitigates the effects of later diversity

Dependent variable:	Voted in 2000		Votes in local/statewide elections	
	Bottom tertile	Top tertile	Bottom tertile	Top tertile
Racial diversity index	0.181	1.235	−0.164	1.058
	(0.248)	(0.439)	(0.158)	(0.337)
Racial SES Gini	−0.006	−0.058	0.064	−0.094
	(0.092)	(0.144)	(0.066)	(0.129)
Share mothers with college degree	0.598	−0.096	0.111	0.038
	(0.299)	(0.284)	(0.274)	(0.267)
Ability grouping	0.053	0.007	−0.064	0.036
	(0.069)	(0.083)	(0.051)	(0.058)
Black	0.167	−0.009	0.097	0.120
	(0.073)	(0.041)	(0.051)	(0.042)
Hispanic	−0.027	−0.043	0.077	0.008
	(0.059)	(0.031)	(0.054)	(0.036)
Asian	0.153	−0.188	−0.174	−0.145
	(0.088)	(0.046)	(0.108)	(0.064)
Other	−0.103	−0.146	0.093	−0.050
	(0.114)	(0.073)	(0.097)	(0.120)
Constant	−0.767	−0.982	0.330	−0.919
	(0.326)	(0.493)	(0.368)	(0.630)
Individual characteristics	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓
Neighborhood characteristics	✓	✓	✓	✓
p-Value, coeffs. equal	0.034		0.001	
N	4,579	3,709	4,035	3,125

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. The omitted category for own race is white. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Second, in Table 1.7, we look at two dependent variables: has voted in 2000 (measured in Wave 3, and one of our main variables of interest in the study), and whether one votes in local/statewide elections (measured in Wave 4). For each, we run the most complete specification on two subsamples: those living in Wave 3 in neighborhoods in the bottom tertile of racial diversity (L, as measured by a neighborhood racial diversity index) and those neighborhoods in the top tertile of racial diversity (H). For the dependent variable measured in Wave 3, we find that the cohort racial diversity index has a positive impact on turnout for the H subgroup—it is significantly higher than for those living in L (p-value = 0.034). For the dependent variable measured in Wave 4, the effects are also in the expected direction, with the cohort racial diversity index having a positive impact on turnout for the H subgroup. Here, it is also significantly higher than in the L subgroup (p-value = 0.001).

We take this as evidence that experiencing racial diversity in adolescence overturns the negative effects on turnout of racial diversity later on, should the results not be due exclusively to selection effects.

Friendships

Evidence in previous sections suggests that peer racial composition matters for voting behavior. The question is: what can explain this long-run positive impact of the racial diversity of one's peers in high school on voting behavior? The literature on racial diversity and voting behavior suggests two likely mechanisms: positive intergroup contact and negative intergroup contact.

On the one hand, according to contact theory, personal contact with outgroup members can reduce prejudice and increase trust under the following conditions (Allport, 1954; Pettigrew and Tropp, 2006; Pettigrew et al., 2011): equal status, shared common goals, a cooperative setting, some form of authority, and friendship potential (summarized by Finseraas et al., 2019). If these conditions are met, exposure to people of different races and social backgrounds in childhood can lead to a more racially diverse friend group later in life. Evidence of such a relation is found by Merlino et al. (2019) for white students, also using the Add Health data set. They show that for whites, being in a cohort with more black individuals of one's own gender has a positive impact on the probability to have a black romantic partner later in life. The authors sug-

gest this is likely due to the higher likelihood of meeting potential partners of other races via friends of one's own gender. These own gender friends are more likely to be black as the share of blacks in the cohort increases. In South Africa, Corno et al. (2022) find that white students assigned to a mixed-race room report a higher share of interracial friendships, have lower prejudice towards blacks, support affirmative action, and are more prosocial in an incentivized experimental game (with a partner of an unspecified race).[32] If positive intergroup experiences predominate, we expect that early interracial contact may reduce the incidence or the magnitude of negative utility from interracial contact later in life. Should this be the case, this increases one's benefits from civic participation later in life (for instance, by voting) in a society that is racially heterogeneous (Alesina and La Ferrara, 2000).

On the other hand, if interracial contact leads to negative experiences, it can stimulate more political activity due to a negative perception of other races. The possibility of ethnic diversity leading to greater prejudice and less trust towards outgroup members is in line with constrict theory (Putnam, 2007).

To determine whether racial diversity in high school leads to more positive or more negative intergroup contact on average, we examine how diversity is linked to two aspects: interracial friendship nominations in Wave 1, and interracial friendships in Wave 4. We use the same econometric specification as when examining voting behavior.

[32]Also for whites, Boisjoly et al. (2006) find that being assigned a black roommate in the first year of university leads to greater openness to minorities, and greater support for affirmative action. However, the study finds no effects on the share of friends of another race, or on the how frequently white students socialize with blacks.

Table 1.8: Friendships

Dependent variable:	Share interracial friendships W1 (1)	White-nonwhite friendship W1 (2)	Has at least one friend of another race W4 (3)
Racial diversity index	0.413	1.210	0.306
	(0.098)	(0.222)	(0.174)
Racial SES Gini	0.011	−0.058	0.016
	(0.029)	(0.080)	(0.054)
Share mothers with college degree	0.084	0.128	0.230
	(0.087)	(0.186)	(0.172)
Ability grouping	−0.011	−0.081	0.004
	(0.022)	(0.044)	(0.043)
Black	0.069	−0.178	−0.011
	(0.021)	(0.046)	(0.029)
Hispanic	0.159	0.107	0.150
	(0.018)	(0.044)	(0.028)
Asian	0.098	0.081	0.243
	(0.031)	(0.066)	(0.028)
Other	0.071	0.112	0.157
	(0.030)	(0.079)	(0.060)
Constant	−0.179	−0.033	−0.477
	(0.133)	(0.342)	(0.260)
Individual characteristics	✓	✓	✓
Family characteristics	✓	✓	✓
Neighborhood characteristics	✓	✓	✓
N	7,802	8,192	10,187

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. The omitted category for own race is white. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

In case the greater political participation in Wave 3 is due to more collaborative and positive contact between members of different racial groups, we can expect a more racially diverse school cohort to be associated with more interracial friendships in Wave 1, while negative experiences with racial diversity in the school cohort could lead to greater racial endogamy. Table 1.8 looks at interracial friendships in Wave 1 in two different ways: in column (1), the dependent variable is the share of friends of other races, as computed by Add Health.[33] Column (2) contains a dummy we constructed from friendship nominations, reflecting whether the respondent has at least one minority friend if the respondent is white and at least one white friend if the respondent is a minority. This variable thus captures interracial friendships which reflect intergroup contact between whites and minorities, rather than among several minorities. In both columns, the coefficients of the racial diversity index are positive and statistically significant. The point estimates imply that an increase of one within-school standard deviation in the index is linked to an increase of 1.3 percentage points in the share of friends of other races in Wave 1 and to an increase of 3.8 percentage points in the probability of a white-minority friendship in Wave 1. This finding is a first indication that negative experiences with racial diversity in school are unlikely to be the dominant explanation for the results in Section IV.A.

Column (3) in Table 1.8 further shows that racial diversity has long-lasting but marginally positive effects on interracial friendships. A survey question in Wave 4 (in 2008, 13–14 years after Wave 1) asked respondents what the races of their close friends were, with the following potential answers: all the same race as myself (1), almost all the same race as myself (2), mostly the same race as myself (3), about half the same race as myself (4), mostly other races than my own (5), almost all other races than my own (6), and all other races than my own (7). Based on these answers, we constructed a dummy variable indicating whether the respondents have close friends of other races (answers (2) to (7) are coded one, and answer (1) is coded zero). Column (3) shows the results of a linear regression that uses this dummy as a dependent variable. We find a marginally positive coefficient, indicating that racial diversity positively impacts

[33]This share is calculated based on all in-school friendship nominations and is not restricted to a student's own school cohort.

the likelihood of people having at least one friend of another race more than a decade later. The point estimate implies that an increase in the racial diversity index of one within-school standard deviation increases this probability by 0.9 percentage points.[34]

Personality

Socialization in early life influences one's personality, preferences, group identities, and beliefs about politics (Madestam and Yanagizawa-Drott, 2012; Algan et al., 2019; Kaplan et al., 2019; Akbulut-Yuksel et al., 2020; Bergman, 2020; Billings et al., 2021). If racial diversity in one's school cohort shapes early life socialization, this constitutes another potential channel through which diversity can impact voting behavior and political preferences of adults.

[34]Coefficients of interest and significance levels are stable if we use weights for longitudinal analyses with Waves 1, 3 and 4.

Table 1.9: Personality in Wave 4

	Average score				
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Imagination/Intellect
Racial diversity index	0.533	−0.059	0.418	−0.098	0.076
	(0.308)	(0.297)	(0.230)	(0.223)	(0.192)
Racial SES Gini	0.019	−0.103	−0.204	0.056	−0.129
	(0.099)	(0.090)	(0.104)	(0.082)	(0.066)
Share mothers with college degree	0.838	0.562	0.176	−0.517	0.209
	(0.313)	(0.180)	(0.296)	(0.257)	(0.217)
	Principal component analysis: score for component 1				
	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Imagination/Intellect
Racial diversity index	1.049	−0.169	0.890	−0.194	0.181
	(0.599)	(0.728)	(0.476)	(0.449)	(0.449)
Racial SES Gini	0.037	−0.256	−0.407	0.085	−0.304
	(0.192)	(0.222)	(0.210)	(0.162)	(0.149)
Share mothers with college degree	1.589	1.325	0.384	−1.083	0.470
	(0.609)	(0.433)	(0.618)	(0.478)	(0.504)
N	10,419	10,424	10,427	10,425	10,349

Notes: All regressions include both cohort composition variables along with controls for cohort fixed effects, school fixed effects, and school trends, as well as the individual student covariates related to the cohort variables. All dependent variables are measured using Wave 4 of the Add Health. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. Wave 4 longitudinal weights are used, for participants who were also interviewed at Waves 1, 3 and 4. Figures in parentheses are standard errors robust to clustering at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

The Add Health survey does not collect data on beliefs or preferences, but it includes a module measuring personality traits using a 20-item mini-IPIP scale for the first time in Wave 4. In Table 1.9, we look at whether personality traits in Wave 4 are influenced by a student's cohort composition. The dependent variables in the two panels in Table 1.9 aggregate the score for each personality trait in two different ways: in the top panel, the average score for all four questions for one trait is used, while in the bottom one, the score for the first component from a principal component analysis for each trait is used. Higher racial diversity leads to marginally more extraversion and conscientiousness, according to both panels. In the estimation sample, all five personality traits measured in Wave 4 are significantly correlated with being registered to vote, with all but neuroticism being negatively correlated, and conscientiousness being only marginally positively correlated. Imagination/intellect is also significantly positively correlated with having voted in 2000 and extraversion is marginally positively so (results from regressions of turnout variables on personality traits and a battery of control variables are available on request). These results are consistent with an earlier study of Cooper et al. (2012), who find that extraversion and conscientiousness are positively related to the probability to be registered to vote.[35]

Results in Table 1.9 provide suggestive evidence that racial diversity in school could have an impact on turnout and political preferences through its shaping of the socialization environment in school. These results are only tentative, and more suitable data is needed to parse out mechanisms more precisely.

IV.D. Heterogeneous effects

Until now, we have seen that racial diversity in a school cohort has long-term positive effects on voting behavior and that racial diversity stimulates friendships with individuals of other races. These findings suggest that, consistent with contact theory, more collaborative and positive contact between individuals of different races could be a driving force behind the increased probability to vote through its impact on early socialization.

[35]The political psychology literature linking personality traits to voter turnout is however best described as having mixed results (Mondak et al., 2010; Mondak, 2010; Gerber et al., 2011a,b).

However, previous research has also shown that the characteristics of individuals engaging in social interactions, such as their race or SES, can influence how they experience the racial composition of their environment (Marschall and Stolle, 2004). In this section, we therefore investigate whether the results on voting behavior differ significantly by racial background or by Wave 1 family income level. We split the sample by minority status (white or minority) and annual family income in Wave 1 (above or below the median of $40,000 in our sample).

Table 1.10: Sample splits: Voting behavior

Dependent variable:	Registered to vote				Voted in 2000			
	White	Minority	Family income $>$ 40k	Family income $\leq$ 40k	White	Minority	Family income $>$ 40k	Family income $\leq$ 40k
Racial diversity index	0.586	0.073	0.645	0.480	0.329	0.633	0.418	0.350
	(0.167)	(0.365)	(0.168)	(0.266)	(0.195)	(0.384)	(0.210)	(0.273)
Racial SES Gini	−0.075	0.051	−0.063	0.026	−0.096	0.051	−0.118	0.017
	(0.092)	(0.126)	(0.071)	(0.106)	(0.069)	(0.103)	(0.096)	(0.063)
Share mothers with college degree	0.066	0.277	0.222	−0.013	0.206	0.179	−0.144	0.627
	(0.162)	(0.305)	(0.186)	(0.195)	(0.206)	(0.235)	(0.223)	(0.236)
Ability grouping	0.003	0.080	−0.047	0.049	0.108	0.057	0.041	0.156
	(0.027)	(0.088)	(0.021)	(0.065)	(0.037)	(0.085)	(0.031)	(0.084)
Black			0.040	0.060			0.105	0.065
			(0.029)	(0.032)			(0.037)	(0.034)
Hispanic		−0.055	0.012	0.025		−0.086	−0.041	−0.029
		(0.034)	(0.032)	(0.031)		(0.035)	(0.038)	(0.032)
Asian		−0.122	−0.146	−0.007		−0.185	−0.133	−0.133
		(0.049)	(0.047)	(0.059)		(0.035)	(0.040)	(0.058)
Other		−0.034	−0.057	0.105		−0.129	−0.103	−0.040
		(0.055)	(0.073)	(0.075)		(0.055)	(0.067)	(0.081)
Constant	−0.456	−1.467	−0.579	−1.116	−1.437	−1.216	−0.855	−1.762
	(0.290)	(0.527)	(0.312)	(0.337)	(0.300)	(0.538)	(0.365)	(0.386)
Individual characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Neighborhood characteristics	✓	✓	✓	✓	✓	✓	✓	✓
p-Value, coeffs. equal	0.202		0.58		0.502		0.835	
N	6,692	5,378	5,947	6,123	6,692	5,378	5,947	6,123

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white in the family income sample splits, and it is black in the minority subsample. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level.All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.11: Sample splits: Political partisanship

Dependent variable:	Identify with a party				Identify as a Democrat			
	White	Minority	Family income $>$ 40k	Family income $\leq$ 40k	White	Minority	Family income $>$ 40k	Family income $\leq$ 40k
Racial diversity index	−0.269	0.131	−0.297	0.228	−0.228	0.082	−0.191	0.066
	(0.206)	(0.333)	(0.339)	(0.212)	(0.150)	(0.319)	(0.223)	(0.147)
Racial SES Gini	0.070	−0.042	0.018	0.123	0.155	−0.009	0.115	0.105
	(0.077)	(0.084)	(0.104)	(0.068)	(0.050)	(0.075)	(0.078)	(0.050)
Share mothers with college degree	0.377	−0.129	0.167	0.185	0.219	−0.368	0.105	0.046
	(0.227)	(0.389)	(0.256)	(0.266)	(0.136)	(0.281)	(0.199)	(0.154)
Ability grouping	−0.050	0.077	−0.143	0.126	0.005	0.027	−0.028	0.068
	(0.054)	(0.084)	(0.071)	(0.038)	(0.042)	(0.093)	(0.056)	(0.032)
Black			0.138	0.134			0.270	0.216
			(0.037)	(0.032)			(0.040)	(0.029)
Hispanic		−0.151	−0.075	−0.002		−0.214	−0.021	0.058
		(0.037)	(0.041)	(0.031)		(0.035)	(0.034)	(0.026)
Asian		−0.247	−0.167	−0.051		−0.293	−0.056	0.016
		(0.045)	(0.038)	(0.059)		(0.045)	(0.039)	(0.057)
Other		−0.166	−0.076	0.061		−0.210	0.004	0.095
		(0.069)	(0.070)	(0.089)		(0.062)	(0.053)	(0.062)
Constant	−0.624	−0.455	−0.727	−0.577	−0.505	0.022	−0.887	−0.121
	(0.307)	(0.396)	(0.348)	(0.285)	(0.197)	(0.412)	(0.271)	(0.219)
Individual characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓	✓
Neighborhood characteristics	✓	✓	✓	✓	✓	✓	✓	✓
p-Value, coeffs. equal	0.293		0.201		0.375		0.324	
N	6,692	5,378	5,947	6,123	6,692	5,378	5,947	6,123

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white in the family income sample splits, and it is black in the minority subsample. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Tables 1.10 and 1.11 present the results for turnout and political preferences, respectively. We find no statistically significant differences between groups, although this is potentially due to smaller sample sizes. However, the fact that there is no clear pattern in the coefficients seems to suggest a positive impact of racial diversity on the voting behavior of both whites and minorities, as well as of individuals from families with an income above or below the median.[36]

V. Robustness checks

V.A. Robustness to attrition and weighting

A potential issue for interpreting our results is that the relationship we find between being registered to vote or having voted in the previous presidential elections and cohort racial diversity might be due to differential attrition in Wave 3. We use two different strategies to check this.[37]

First, we estimate equation (1.1) using as dependent variables indicator variables for being a respondent in Waves 2, 3 and 4 (as Wave 1 was not affected by attrition). For all three dependent variables, in all three specifications (including individual characteristics, then gradually adding family characteristics and neighborhood characteristics) the coefficients of the racial diversity index are close to zero and insignificant.[38]

Second, in our estimations we use Wave 3 longitudinal weights. This places more weight on those in categories from which there are more attriters. Our results may be due to some observations being weighted more heavily. To check this, we drop from the estimation sample those respondents with the highest 10% Wave 3 longitudinal weights. We re-estimate the regressions in Table 1.4 using this smaller sample. Results are presented in Appendix Table 1.19. The coefficients of the racial diversity index decrease, but they remain significant at the 1% level for being registered to vote and at 10% level for having voted in 2000. In this sample, the racial SES Gini has a negative and marginally signifi-

[36] We also run separate analyses for turnout by region (West, Midwest, Northeast, South) and by cohort size (above or below the median). While the coefficients of the racial diversity index are larger in some subsamples than in others, none is significantly different from its counterparts. These results are available in Web Appendix Tables 15 and 16.

[37] The procedures in this Section have been inspired by those in Merlino et al. (2019).

[38] Results available in the Web Appendix Tables 21–23.

cant coefficient in the regression with having voted as a dependent variable.[39]

From these tests, we conclude that attrition does not affect our estimates significantly.

V.B. Robustness to different specifications of racial diversity

In Appendix Tables 1.17 and 1.18, we test the robustness of our results to other specifications of racial diversity. We add the racial shares, use only the racial shares, or use the racial shares plus the racial shares squared. The coefficient of the racial diversity index remains positive and significant for the probability to be registered to vote (column (1) in Table 1.17), and remains positive but insignificant for the probability to have voted in 2000 (column (4) in Table 1.17).

For voting behavior (Table 1.17), the share of Asians is significant, regardless of whether we include the squared shares or not. For political partisanship (Table 1.18), a higher share of black students in the cohort is positively related to respondents identifying with a political party, particularly with the Democratic Party (though this is not significant in all specifications). In all specifications, more interracial SES inquality (as measured by the racial SES Gini) significantly increases the probability that one identifies as a Democrat. Following Madestam and Yanagizawa-Drott (2012), we interpret this as suggesting that greater racial inequality in one's cohort shifts preferences to the political left without increasing political polarization (since it has no effect on the probability to identify with the Republican Party—results confirming this are available

[39]We also trimmed the sample manually using Lee bounds (Lee, 2009). This procedure provides bounds for the treatment effect, under the assumption that the effect of the treatment on attrition is monotonic—in our case, that those in more racially diverse cohorts as measured by a higher racial diversity index are more likely to attrit by Wave 3 than those in less racially diverse cohorts. We define a cohort as treated if its racial diversity index is above the mean of the racial diversity indices of all cohorts in the school.

In practice, the method drops observations (from either the Wave 1 in-home sample or from the Wave 3 sample) in a way that equalizes the share of treated in the in-home Wave 1 sample and in the Wave 3 sample. It drops one of two types of observations: either those that contribute the most to the correlation between the treated dummy and the dependent variable (Registered to vote, Voted in 2000)—which gives a lower bound for the treatment effect, or the observations that contribute the least to the correlation between the treated dummy and the dependent variable—which gives an upper bound for the treatment effect. We manually select the observations to be dropped. For both turnout variables, the lower bound is positive, indicating that our results are not significantly biased upward due to attrition.

upon request).

These results are in line with Kaplan et al. (2019); Bergman (2020); Billings et al. (2021). These papers find that a higher share of minorities in one's school (or a proxy for it, namely assignment to busing or the option to transfer to a more racially diverse school than one's initial school) has no effect on turnout and registration. All three papers find that the specific desegregation policy they analyze makes their respondents more likely to identify as Democrats: Kaplan et al. (2019) finds this effect for white males, Bergman (2020) for minority students, and Billings et al. (2021) only for white students.

V.C. Potential issues with self-reported turnout

The literature on voter turnout acknowledges that self-reported turnout is consistently higher than actual turnout (for instance, see Enamorado and Imai, 2019). This is also most likely the case with the two turnout variables used in the analysis, Registered to Vote and Voted in 2000. For instance, administrative data for the November 2000 elections for 18—24 year-olds (the age group to which 94.75% of the estimation sample belongs in Wave 3, which was conducted between 10 and 17 months after the elections) reports 45.4% were registered to vote (compared to 73.8% self-reported registration in the estimation sample). The official turnout was 32.3% in this age group (compared to 44.2% self-reported turnout in the estimation sample, see Table 1.1).[40]

This could affect our results if overreporting turnout is systematically correlated with racial diversity at the cohort level. For instance, the treatment (being in a more racially diverse cohort than the school average) could increase social desirability bias, which in turn could increase self-reported voting.

We check whether our estimates might be driven by social desirability bias in several ways. If, for instance, individuals in more racially diverse cohorts are more agreeable or more likely to think it is important to fit in with one's group, then our estimates of voter registration and turnout might be biased upwards (since the dependent variables are self-reported). The cohort racial diversity has no impact on agreeableness in Wave 4 (see column 2 in both panels in Table 1.9). We also estimate equation (1.1) using as dependent variable a

[40]Source for the administrative data: `https://www.census.gov/data/tables/2000/demo/voting-and-registration/p20-542.html`, accessed on September 6, 2022.

binary variable from Wave 3, reflecting the perceived importance to fit in with one's group. We coded responses of "agree" and "strongly agree" as 1 and the rest as 0.

The coefficient of the racial diversity index in the most complete specification (including individual characteristics, family characteristics, neighborhood characteristics, plus grade fixed effects, school fixed effects and school linear trends) is -0.153, and is insignificant (the standard error is 0.132). These results suggest that social desirability bias, as captured by the desire to fit in and by agreeableness, does not impact our results significantly.

V.D. Relating our results to the literature

We find positive long-term effects of racial diversity in adolescence on voter registration and turnout. Is this surprising, given the negative or inconsistent short-term effects of racial diversity found in other studies?

In order to answer this question, we compare the effects of the racial diversity index on other short-term and long-term behaviors which have been investigated by previous studies. Appendix Tables 1.20 and 1.21 present the relationship between the racial diversity index and Wave 1 and Wave 3 behaviors which have been found to be sensitive to peer influence. The regressions use our most comprehensive specification and show that even if there is some evidence of a negative short-term correlation between the racial diversity index and behavior in Wave 1, this mostly vanishes by Wave 3. This is similar to results in Bifulco et al. (2011) and Bifulco et al. (2014), who find no long-term effects of the share of minorities (the cumulative shares of blacks and Hispanics) in one's cohort on post-secondary outcomes. We also observe that racial diversity has a significant negative impact on binge drinking and a marginally positive impact on test scores later in life.

These results indicate that our main findings are in line with previous findings. The most plausible reason for any apparent difference is therefore the different time frame for the effects (long- versus short-run), the different geographic aggregation of the data (narrowly versus broadly defined peer groups), or using a different measure of racial diversity.

V.E. How likely is it to find any long-term effect of the racial diversity index?

In Section V.D we checked whether there are effects of the racial diversity index on several long-term outcomes. This raises the issue that the more tests we run, the higher the chance that some results turn out significant. We thus test the composite null of no effects of racial diversity in the long term on the following variables: our main variables of interest (is registered to vote, has voted in 2000) plus the seven variables from Section V.D (is a high school dropout, has a college degree, the score in the Picture Vocabulary Test administered by Add Health, is idle (does not work and does not attend school), smokes, smokes marijuana, engages in binge drinking).

We use a resampling procedure, as discussed in Bifulco et al. (2011, section I.C). The authors combine the resampling approach by Westfall and Young (1993, p. 214–215) with a strategy by Agresti (2002, p. 97–98) to calculate the likelihood that a certain pattern of *p*-values might arise should the composite null hypothesis that there are no effects of racial cohort composition be true. The probability of a false positive is the sum of the probabilities of all possible outcomes that occur with a probability lower than or equal to the probability in the observed data. The share of *p*-values corresponding to outcomes more extreme than those in the observed data is calculated using this resampling approach.

We estimate a linear-in-means model for the nine Wave 3 outcome variables. We regress these variables on the battery of characteristics used in our main specification, excluding the racial diversity index and the racial SES Gini index, but including school fixed effects, cohort fixed effects, either with or without school linear trends. We run 10,000 simulations following the procedure in Bifulco et al. (2011). For the racial diversity index, values more extreme than the one observed under the null hypothesis are quite unlikely (p-value = 0.002 without trends, p-value = 0.004 with trends).[41]

[41] Results are available on request.

V.F. Are cross-cohort spillovers a matter of concern?

We considered the possibility that our results might suffer from cross-cohort spillovers, as pupils also interact with others in other cohorts. While this is true, most respondents have most of their friends in their own grade. In the school survey in Wave 1, pupils were asked to nominate up to 5 female and 5 male friends. They could nominate anyone, also people from other schools. From the friends in their school (mean: 73%, median: 85% of all friends),[42] in-school respondents mostly nominate friends from their grade (mean: 75%, median: 86% of the friends whose grade could be retrieved). 45% of respondents in this sample nominate exclusively friends from their own school cohort. From this, we conclude that if there exist cross-cohort spillovers, their impact is likely to be negligible relative to the within-cohort effects.

VI. Discussion

This paper finds that racial diversity in high school has a positive impact on individuals' voting behavior in early adulthood. We show that this result is likely due to positive and persistent interracial contact. Respondents exposed to more quasi-randomly occurring diversity in adolescence have more friends of other races, both in high school and more than a decade later. The point estimates suggest that the effect sizes are nontrivial when compared to "get out the vote" initiatives, which increase turnout by 0.2–0.3 percentage points (the possibility to be asked about one's voting experience in a phone survey, DellaVigna et al., 2016; Rogers et al., 2016) up to 7.1 percentage points (door-to-door canvassing, de Rooij et al., 2009). An increase of one within-school standard deviation in the racial diversity index leads to an increase of 1.7 percentage points in the probability to be registered to vote seven years later and an increase of 1.1 percentage points in the probability to have voted six years later.

These results underscore that beyond their instrumental role as transmitters of knowledge, schools are important arenas for socialization, a role that is often overlooked by research (Gradstein and Justman, 2000, 2002). This role should be considered in the design of educational policies with a focus on

[42]The difference between the mean and the median is due to 13.7% of respondents skipping the friendship nomination section in the survey or not nominating any friend.

racial diversity. However, more research is needed to understand whether our results are generalizable to other contexts, especially to contexts in which increases in diversity are of a greater order of magnitude than those studied in this paper, or to contexts in which diversity is imposed exogenously rather than arising by chance. Another direction for future research is to better understand the channels—such as beliefs or preferences—through which early intergroup contact affects voting in the long run.

Appendix

1.A. Definitions, sample restrictions, and other descriptives

Table 1.12: Description of variables

Variable	Wave	Description	Values
Dependent variables			
Registered to vote	3	Reports being registered to vote	No = 0, Yes = 1
Voted in 2000	3	Reports having voted	No = 0, Yes = 1
Identifies with a party	3	Reports identifying with a political party	No = 0, Yes = 1
– Democrat	3	"Yes" to previous question & reports identifying with the Democratic Party	Other/No = 0, Yes = 1
– Republican	3	"Yes" to previous question & reports identifying with the Republican Party	Other/No = 0, Yes = 1
School cohort composition variables			
Share males in cohort	1	Share of male students in one's cohort	[0,1]
Share black/Hispanic/Asian/other in cohort	1	Share of students in an individual's cohort who define themselves to be black/Hispanic/Asian/other (omitted: white)	[0,1]
Share mothers with college degree	1	Share of students in an individual's cohort whose mothers have a college degree	[0,1]
Racial diversity index	1	One minus the sum of squared racial shares in one's cohort	[0,0.8]
Racial SES Gini	1	Definition in footnote 12	[0,1]
Family characteristics			
Mother's education	1	Dummies for high school dropout, high school graduate, some college, college graduate (imputed if missing)	
Family income	1	Imputed annual family income of individual (log in regression)	in 000's. USD
English spoken at home	1	Dummy variable	No = 0, Yes = 1
Lives with both biological parents	1	Dummy variable	No = 0, Yes = 1
Parent civically engaged	1	Dummy variable if parent answering is a member of any of the following: parent/teacher organization, military veterans organization, labor union, hobby/sports group, civic or social organization (imputed if missing)	No = 0, Yes = 1
Parent dummy	1	Dummy variable if missing parent information on either mother's education, family income, parent's age, parent's civic engagement	No = 0, Yes = 1

Variable	Wave	Description	Values
Neighborhood characteristics			
Urban area	1	Respondent lives in urban area	No = 0, Yes = 1
Share of census block group[a] with less than high school education	1		[0,1]
Share of census block group with a bachelor's degree	1		[0,1]
Share votes for Democratic candidate 1992	1		[0,1]
Share of census block group black/Hispanic/Asian & other	1		[0,1]
Share of census block group below poverty level	1	Share of inhabitants in the census block group with income in 1989 below poverty level	[0,1]
Personality			
Extraversion	4	Average of 4 items (or first principal component) in 20-item mini IPIP scale	[1,5]
Agreeableness	4	Average of 4 items (or first principal component) in 20-item mini IPIP scale	[1,5]
Conscientiousness	4	Average of 4 items (or first principal component) in 20-item mini IPIP scale	[1,5]
Neuroticism	4	Average of 4 items (or first principal component) in 20-item mini IPIP scale	[1,5]
Imagination/Intellect	4	Average of 4 items (or first principal component) in 20-item mini IPIP scale	[1,5]
Other variables			
Ability grouping	1	Principal answered whether English or language arts classes in a grade are grouped by ability	No = 0, Yes = 1
Share interracial friendships	1	Share of friends of other races, as computed by Add Health	[0,1]
White-nonwhite friendship	1	At least a white/minority friend, if minority/white	No = 0, Yes = 1
Has at least a friend of another race	4	Dummy variable	No = 0, Yes = 1
Important to fit in	3	(Strongly) agreed it is important to fit in with group	No = 0, Yes = 1
Drop out of high school	3	Dummy variable	No = 0, Yes = 1
College degree	3	Dummy variable	No = 0, Yes = 1
Test score	1, 3	Standardized Add Health picture vocabulary test score	[9,123]
Idleness	3	Not in school and not working	No = 0, Yes = 1
Smoking	1, 3	Smoked in the past 30 days	No = 0, Yes = 1
Marijuana use	1, 3	Used in the past 30 days	No = 0, Yes = 1
Binge drinking	1, 3	Had drunk at least 5 drinks in a row at least once in the past 12 months	No = 0, Yes = 1

[a] A *census block group* is a cluster of census blocks within a census tract or block numbering area. It is the lowest geographical level for which the Census Bureau publishes sample data. For the 1990 census, block groups averaged 452 housing units, or 1,100 people. A typical census tract contains 4 or 5 block groups.

Table 1.13: Sample restrictions

Wave 1 respondents followed in Wave 3	14,979
Longitudinal weights	14,322
Racial diversity index can be computed	13,501
American citizens in November 2000	12,964
Turnout variables in Wave 3	12,821
Political preferences in Wave 3	12,677
Individual characteristics	12,558
Family characteristics	12,558
Neighborhood characteristics in Wave 1	12,430
Cohort controls	12,206
At least 10 pupils in cohort	12,150
School has at least two cohorts	12,070

Notes: The table shows how each sample restriction contributes to the size of the estimation sample. Each row shows the number of respondents who have valid responses for the respective variable(s), as well as for all the variables listed above it. Individual, family and neighborhood characteristics are listed in Appendix Table 1.12.

Table 1.14: Summary statistics: other variables

	All		White		Minority	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Wave 1: individual characteristics						
Male	0.508	0.500	0.505	0.500	0.514	0.500
Age	15.902	1.793	15.854	1.782	16.010	1.813
GPA	2.803	0.774	2.870	0.786	2.655	0.723
Wave 1: family characteristics						
Mother high school dropout	0.156	0.363	0.112	0.315	0.255	0.436
Mother high school graduate	0.377	0.485	0.387	0.487	0.356	0.479
Mother some college	0.230	0.421	0.246	0.431	0.194	0.395
Mother college graduate	0.237	0.425	0.256	0.436	0.196	0.397
Family income (k)	45.472	38.932	50.012	38.846	35.301	37.166
English not spoken at home	0.051	0.220	0.004	0.063	0.156	0.363
Lives with both biological parents	0.572	0.465	0.617	0.461	0.472	0.457
Parent civically engaged	0.510	0.472	0.549	0.474	0.422	0.454
Missing parent information	0.306	0.461	0.266	0.442	0.395	0.489
Wave 1: neighborhood characteristics						
Share less than high school	0.272	0.154	0.242	0.135	0.337	0.173
Share with bachelor's degree	0.222	0.144	0.231	0.145	0.201	0.139
Share votes for Democratic candidate 1992	0.423	0.095	0.405	0.084	0.462	0.107
Share blacks	0.138	0.261	0.048	0.114	0.340	0.365
Share Hispanics	0.062	0.142	0.029	0.061	0.136	0.221
Share Asians & other	0.031	0.081	0.018	0.043	0.060	0.127
Share below poverty level	0.142	0.138	0.110	0.104	0.213	0.173
Urban area	0.499	0.500	0.427	0.495	0.660	0.474
N	12,070		6,692		5,378	

Notes: Summary statistics are calculated using Wave 3 longitudinal weights, which aim to produce a representative sample of individuals who were surveyed in both Waves 1 and 3.

	All		White		Minority	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Wave 1: cohort controls						
Share mothers with college degree	0.296	0.139	0.296	0.138	0.298	0.143
Share males	0.499	0.072	0.503	0.077	0.489	0.057
Ability grouping	0.456	0.498	0.436	0.496	0.500	0.500
Wave 1: other						
Average class size in school	25.691	4.646	24.957	4.378	27.334	4.805
Share interracial friendships	0.231	0.225	0.196	0.209	0.324	0.238
White-non white friend	0.466	0.499	0.504	0.500	0.366	0.482
Region						
Northeast	0.138	0.344	0.153	0.360	0.104	0.305
Midwest	0.301	0.459	0.358	0.479	0.173	0.379
South	0.403	0.490	0.361	0.480	0.498	0.500
West	0.159	0.365	0.129	0.335	0.225	0.418
Wave 3 variables						
Married	0.163	0.369	0.178	0.383	0.129	0.335
Education (years)	13.133	1.979	13.230	1.976	12.914	1.967
Working	0.745	0.436	0.762	0.426	0.705	0.456
Annual income (k)	13.775	14.978	14.233	14.172	12.707	16.659
It is important to fit in	0.341	0.474	0.374	0.484	0.267	0.442
N	12,070		6,692		5,378	

Notes: "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. Summary statistics are calculated using Wave 3 longitudinal weights, which aim to produce a representative sample of individuals who were surveyed in both Waves 1 and 3.

The variables "Share interracial friendships" and "White-non white friend" have fewer responses: 7,802 (3,275 white and 4,527 minority respondents), respectively 8,192 (3,455 white and 4,737 minority respondents).

	All		White		Minority	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Wave 1: behavior						
Smoking	0.264	0.441	0.303	0.460	0.177	0.381
Marijuana use	0.139	0.346	0.134	0.341	0.149	0.357
Binge drinking	0.270	0.444	0.293	0.455	0.219	0.414
Wave 3: behavior						
Drop out of high school	0.171	0.377	0.155	0.362	0.207	0.405
Attend college	0.567	0.495	0.594	0.491	0.508	0.500
Test score	102.208	14.369	105.149	11.420	95.619	17.722
Idleness	0.133	0.339	0.114	0.318	0.175	0.380
Smoking	0.360	0.480	0.409	0.492	0.248	0.432
Marijuana use	0.232	0.422	0.247	0.431	0.199	0.400
Binge drinking	0.517	0.500	0.590	0.492	0.353	0.478
Wave 4 variables						
Extraversion	3.310	0.774	3.341	0.779	3.237	0.759
Agreeableness	3.812	0.607	3.837	0.608	3.754	0.600
Conscientiousness	3.642	0.673	3.625	0.691	3.681	0.627
Neuroticism	2.596	0.686	2.570	0.690	2.657	0.673
Imagination/Intellect	3.628	0.621	3.648	0.628	3.580	0.599
Has at least one friend of another race	0.540	0.498	0.512	0.500	0.608	0.488
N	12,070		6,692		5,378	

Notes: Summary statistics are calculated using Wave 3 longitudinal weights for variables in Wave 1, which aim to produce a representative sample of individuals who were surveyed in both Waves 1 and 3. For variables in Wave 4, we use weights for longitudinal analyses with Waves 1, 3 and 4. Personality trait scores are calculated as averages over 4 questions with answers ranging from 1 to 5, where higher numbers indicate more of that trait than lower numbers.
The variable "Has at least one friend of another race" has fewer responses: 10,187 (4,358 white and 5,829 minority respondents).

1.B. Robustness checks

Table 1.15: Robustness check: Restricted sample (Lavy et al., 2012)

Dependent variable:	Registered to vote			Voted in 2000		
	(1)	(2)	(3)	(4)	(5)	(6)
Racial diversity index	0.586	0.579	0.578	0.364	0.359	0.356
	(0.170)	(0.167)	(0.168)	(0.227)	(0.223)	(0.223)
Racial SES Gini	−0.049	−0.053	−0.052	−0.045	−0.045	−0.045
	(0.080)	(0.077)	(0.078)	(0.062)	(0.060)	(0.061)
Share mothers with college degree	−0.025	−0.041	−0.044	0.514	0.460	0.461
	(0.204)	(0.209)	(0.208)	(0.223)	(0.218)	(0.221)
Ability grouping	0.001	0.004	0.001	0.068	0.070	0.069
	(0.035)	(0.031)	(0.033)	(0.042)	(0.040)	(0.041)
Black	0.045	0.061	0.051	0.060	0.090	0.071
	(0.019)	(0.020)	(0.022)	(0.024)	(0.024)	(0.022)
Hispanic	−0.034	0.007	0.017	−0.082	−0.037	−0.026
	(0.025)	(0.023)	(0.024)	(0.020)	(0.019)	(0.020)
Asian	−0.127	−0.098	−0.095	−0.159	−0.133	−0.134
	(0.039)	(0.038)	(0.038)	(0.036)	(0.029)	(0.029)
Other	−0.035	−0.009	−0.012	−0.121	−0.094	−0.102
	(0.051)	(0.050)	(0.052)	(0.052)	(0.053)	(0.054)
Constant	−0.695	−0.809	−0.840	−1.397	−1.544	−1.457
	(0.250)	(0.244)	(0.269)	(0.304)	(0.296)	(0.311)
Individual characteristics	✓	✓	✓	✓	✓	✓
Family characteristics		✓	✓		✓	✓
Neighborhood characteristics			✓			✓
N	10,960	10,960	10,960	10,960	10,960	10,960

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.16: Robustness check: Restricted sample (Bifulco et al., 2011)

Dependent variable:	Registered to vote			Voted in 2000		
	(1)	(2)	(3)	(4)	(5)	(6)
Racial diversity index	0.562	0.555	0.553	0.503	0.496	0.492
	(0.164)	(0.163)	(0.165)	(0.149)	(0.142)	(0.144)
Racial SES Gini	−0.029	−0.038	−0.039	−0.011	−0.014	−0.014
	(0.082)	(0.079)	(0.079)	(0.049)	(0.046)	(0.047)
Share mothers with college degree	0.117	0.078	0.078	0.308	0.244	0.245
	(0.149)	(0.148)	(0.150)	(0.167)	(0.165)	(0.166)
Ability grouping	0.001	0.005	0.002	0.084	0.087	0.085
	(0.031)	(0.028)	(0.029)	(0.036)	(0.036)	(0.037)
Black	0.053	0.069	0.058	0.060	0.089	0.062
	(0.019)	(0.019)	(0.021)	(0.024)	(0.023)	(0.022)
Hispanic	−0.041	0.002	0.012	−0.087	−0.045	−0.033
	(0.024)	(0.022)	(0.023)	(0.020)	(0.020)	(0.020)
Asian	−0.131	−0.098	−0.093	−0.159	−0.136	−0.136
	(0.037)	(0.036)	(0.036)	(0.035)	(0.029)	(0.029)
Other	−0.034	−0.004	−0.004	−0.119	−0.092	−0.097
	(0.048)	(0.047)	(0.049)	(0.049)	(0.050)	(0.052)
Constant	−0.673	−0.769	−0.823	−1.208	−1.330	−1.257
	(0.237)	(0.235)	(0.257)	(0.260)	(0.250)	(0.272)
Individual characteristics	✓	✓	✓	✓	✓	✓
Family characteristics		✓	✓		✓	✓
Neighborhood characteristics			✓			✓
N	11,538	11,538	11,538	11,538	11,538	11,538

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.17: Robustness check: Alternative specification, voting behavior

Dependent variable:	Registered to vote			Voted in 2000		
	(1)	(2)	(3)	(4)	(5)	(6)
Racial diversity index	0.509			0.182		
	(0.204)			(0.237)		
Racial SES Gini	−0.039	−0.033	−0.028	−0.063	−0.061	−0.058
	(0.075)	(0.075)	(0.074)	(0.050)	(0.050)	(0.052)
Share mothers with college degree	0.098	0.109	0.084	0.194	0.198	0.196
	(0.143)	(0.145)	(0.145)	(0.162)	(0.164)	(0.162)
Ability grouping	−0.002	−0.005	0.006	0.080	0.079	0.084
	(0.029)	(0.029)	(0.030)	(0.040)	(0.040)	(0.041)
Share black	−0.107	0.152	−0.144	0.080	0.173	0.029
	(0.268)	(0.271)	(0.668)	(0.238)	(0.266)	(0.716)
Share Hispanic	0.086	0.624	1.452	0.388	0.580	0.875
	(0.426)	(0.476)	(0.512)	(0.361)	(0.410)	(0.565)
Share Asian	0.502	0.938	1.576	0.909	1.065	1.513
	(0.326)	(0.350)	(0.655)	(0.430)	(0.392)	(0.687)
Share other	−0.140	0.564	0.647	0.010	0.262	0.446
	(0.486)	(0.373)	(0.819)	(0.550)	(0.389)	(0.726)
Share black squared			0.310			0.165
			(0.783)			(0.790)
Share Hispanic squared			−2.385			−0.884
			(0.603)			(0.742)
Share Asian squared			−1.825			−1.259
			(1.212)			(1.524)
Share other race squared			−1.616			−1.927
			(6.565)			(3.380)
Black	0.059	0.059	0.059	0.064	0.064	0.064
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Hispanic	0.012	0.012	0.011	−0.034	−0.034	−0.034
	(0.023)	(0.023)	(0.023)	(0.020)	(0.020)	(0.020)
Asian	−0.092	−0.091	−0.092	−0.135	−0.135	−0.135
	(0.036)	(0.036)	(0.036)	(0.029)	(0.029)	(0.029)
Other	−0.002	−0.006	−0.001	−0.102	−0.103	−0.101
	(0.049)	(0.049)	(0.050)	(0.051)	(0.051)	(0.051)
Constant	−0.815	−0.741	−0.704	−1.219	−1.192	−1.180
	(0.253)	(0.240)	(0.253)	(0.287)	(0.273)	(0.286)
Individual characteristics	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓
Neighborhood characteristics	✓	✓	✓	✓	✓	✓
N	12,070	12,070	12,070	12,070	12,070	12,070

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.18: Robustness check: Alternative specification, political partisanship

Dependent variable:	Identifies with a party			Identifies as a Democrat		
	(1)	(2)	(3)	(4)	(5)	(6)
Racial diversity index	−0.849			−0.214		
	(0.489)			(0.393)		
Share black	2.698	0.419	1.471	0.883	0.524	0.574
	(0.846)	(0.243)	(0.659)	(0.659)	(0.242)	(0.563)
Share Hispanic	0.558	−0.657	−0.828	−0.461	−0.688	−0.810
	(0.785)	(0.428)	(0.635)	(0.688)	(0.327)	(0.471)
Share Asian	0.903	−0.387	−0.435	0.614	0.227	0.277
	(0.927)	(0.507)	(0.715)	(0.691)	(0.317)	(0.518)
Share other	0.745	−0.163	−0.584	0.423	−0.262	0.088
	(1.012)	(0.395)	(0.680)	(0.951)	(0.278)	(0.543)
Share black squared	−2.900		−1.604	−0.365		−0.039
	(1.006)		(0.784)	(0.821)		(0.724)
Share Hispanic squared	−1.386		0.291	−0.131		0.292
	(1.156)		(0.935)	(0.980)		(0.686)
Share Asian squared	−1.618		0.075	−0.486		−0.059
	(1.636)		(1.295)	(1.181)		(0.942)
Share other race squared	2.035		3.697	−3.108		−2.689
	(4.103)		(4.575)	(2.830)		(2.584)
Racial SES Gini	0.059	0.072	0.058	0.107	0.107	0.107
	(0.052)	(0.053)	(0.052)	(0.032)	(0.031)	(0.032)
Share mothers with college degree	0.185	0.174	0.177	0.027	0.017	0.026
	(0.177)	(0.187)	(0.182)	(0.125)	(0.126)	(0.126)
Ability grouping	−0.052	−0.040	−0.049	−0.004	−0.003	−0.003
	(0.054)	(0.052)	(0.054)	(0.039)	(0.038)	(0.039)
Black	0.114	0.115	0.114	0.213	0.213	0.213
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
Hispanic	−0.033	−0.033	−0.033	0.028	0.028	0.028
	(0.025)	(0.025)	(0.025)	(0.021)	(0.021)	(0.021)
Asian	−0.126	−0.124	−0.125	−0.031	−0.031	−0.031
	(0.034)	(0.033)	(0.033)	(0.035)	(0.035)	(0.035)
Other	−0.012	−0.014	−0.012	0.048	0.049	0.048
	(0.057)	(0.057)	(0.057)	(0.041)	(0.041)	(0.041)
Constant	−0.520	−0.578	−0.619	−0.373	−0.383	−0.398
	(0.250)	(0.244)	(0.245)	(0.215)	(0.203)	(0.205)
Individual characteristics	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓
Neighborhood characteristics	✓	✓	✓	✓	✓	✓
N	12,070	12,070	12,070	12,070	12,070	12,070

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.19: Robustness to attrition and weighting

Dependent variable:	Registered to vote			Voted in 2000		
	(1)	(2)	(3)	(4)	(5)	(6)
Racial diversity index	0.423	0.406	0.409	0.307	0.281	0.276
	(0.148)	(0.150)	(0.151)	(0.150)	(0.145)	(0.145)
Racial SES Gini	0.000	−0.007	−0.007	−0.085	−0.088	−0.089
	(0.052)	(0.052)	(0.052)	(0.053)	(0.052)	(0.052)
Share mothers with college degree	0.337	0.300	0.297	0.115	0.079	0.075
	(0.119)	(0.123)	(0.124)	(0.153)	(0.149)	(0.151)
Ability grouping	0.010	0.014	0.013	0.091	0.092	0.092
	(0.061)	(0.057)	(0.057)	(0.036)	(0.034)	(0.034)
Black	0.042	0.052	0.044	0.047	0.070	0.052
	(0.019)	(0.018)	(0.021)	(0.021)	(0.020)	(0.021)
Hispanic	−0.023	0.019	0.027	−0.059	−0.020	−0.013
	(0.023)	(0.022)	(0.023)	(0.021)	(0.021)	(0.021)
Asian	−0.086	−0.063	−0.060	−0.111	−0.094	−0.097
	(0.032)	(0.030)	(0.031)	(0.038)	(0.034)	(0.034)
Other	−0.035	−0.011	−0.012	−0.131	−0.108	−0.109
	(0.051)	(0.047)	(0.048)	(0.052)	(0.051)	(0.051)
Constant	−0.701	−0.738	−0.770	−0.900	−0.986	−0.796
	(0.224)	(0.220)	(0.252)	(0.253)	(0.244)	(0.264)
Individual characteristics	✓	✓	✓	✓	✓	✓
Family characteristics		✓	✓		✓	✓
Neighborhood characteristics			✓			✓
N	10,863	10,863	10,863	10,863	10,863	10,863

Notes: This table reports OLS estimates. Controls include school and grade fixed effects and school linear trends. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. The omitted category for own race is white. Wave 3 longitudinal weights are used. Standard errors (in parentheses) are clustered at the school level. The sample drops individuals from the estimation sample who have the top 10% highest Wave 3 longitudinal weights. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.20: Behavior in Wave 1

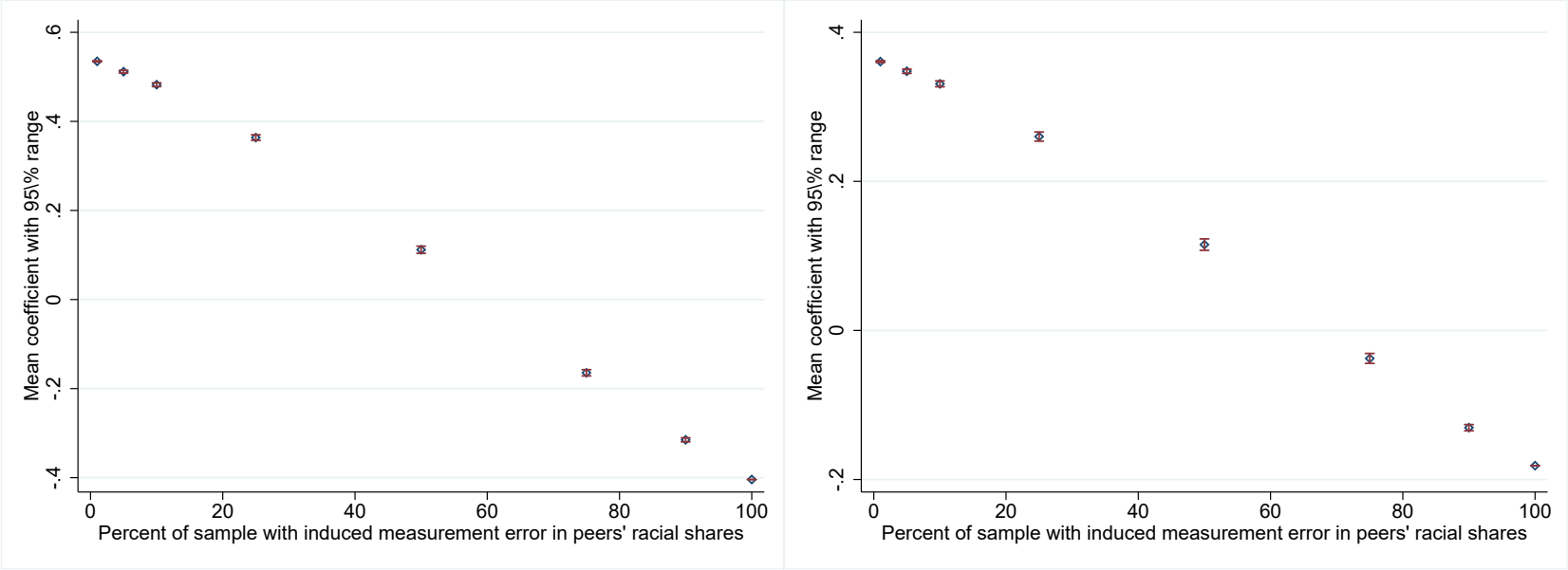
	Smoking	Marijuana use	Binge drinking
Racial diversity index	0.200	0.150	0.323
	(0.189)	(0.085)	(0.124)
Racial SES Gini	0.035	−0.038	−0.099
	(0.057)	(0.039)	(0.046)
Share mothers with college degree	−0.231	−0.240	−0.155
	(0.157)	(0.110)	(0.119)
Ability grouping	0.052	0.029	0.059
	(0.028)	(0.047)	(0.033)
Black	−0.191	−0.005	−0.123
	(0.024)	(0.022)	(0.023)
Hispanic	−0.022	0.036	0.042
	(0.025)	(0.022)	(0.022)
Asian	−0.044	−0.014	−0.077
	(0.028)	(0.028)	(0.026)
Other	−0.079	0.087	−0.091
	(0.038)	(0.045)	(0.029)
Constant	0.514	0.326	0.401
	(0.229)	(0.147)	(0.197)
Individual characteristics	✓	✓	✓
Family characteristics	✓	✓	✓
Neighborhood characteristics	✓	✓	✓
N	12,502	12,391	12,540

Notes: All regressions include both cohort composition variables along with controls for cohort fixed effects, school fixed effects, and school trends, as well as the individual student covariates related to the cohort variables. All dependent variables are measured using Wave 1 of the Add Health. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. Wave 1 cross-sectional weights are used. Figures in parentheses are standard errors robust to clustering at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.

Table 1.21: Behavior in Wave 3

	Drop out of high school	College degree	Test score	Idleness	Smoking	Marijuana use	Binge drinking
Racial diversity index	−0.010	−0.034	5.178	0.120	0.282	0.100	−0.387
	(0.101)	(0.126)	(2.978)	(0.111)	(0.177)	(0.136)	(0.189)
Racial SES Gini	0.018	−0.032	−0.139	−0.004	0.032	0.039	0.160
	(0.033)	(0.053)	(1.093)	(0.042)	(0.063)	(0.053)	(0.073)
Share mothers with college degree	−0.186	0.281	5.502	0.088	0.176	−0.270	0.057
	(0.096)	(0.146)	(3.532)	(0.097)	(0.136)	(0.140)	(0.162)
Ability grouping	0.018	0.008	0.587	0.012	−0.007	0.016	−0.065
	(0.022)	(0.027)	(0.766)	(0.018)	(0.042)	(0.037)	(0.036)
Black	−0.055	0.053	−7.009	0.036	−0.215	−0.029	−0.256
	(0.019)	(0.019)	(0.754)	(0.019)	(0.023)	(0.023)	(0.029)
Hispanic	0.003	0.016	−1.724	0.024	−0.053	0.003	−0.028
	(0.022)	(0.024)	(0.918)	(0.021)	(0.027)	(0.029)	(0.031)
Asian	−0.019	0.086	−2.121	0.015	0.001	−0.047	−0.144
	(0.023)	(0.032)	(0.930)	(0.027)	(0.037)	(0.035)	(0.040)
Other	0.058	0.047	−0.551	0.055	−0.040	0.049	0.043
	(0.050)	(0.035)	(1.815)	(0.052)	(0.045)	(0.058)	(0.051)
Constant	−1.122	1.232	123.549	0.547	0.440	1.179	0.948
	(0.170)	(0.205)	(5.725)	(0.181)	(0.242)	(0.217)	(0.236)
Individual characteristics	✓	✓	✓	✓	✓	✓	✓
Family characteristics	✓	✓	✓	✓	✓	✓	✓
Neighborhood characteristics	✓	✓	✓	✓	✓	✓	✓
N	12,066	11,360	11,657	11,635	12,017	12,033	12,021

Notes: All regressions include both cohort composition variables along with controls for cohort fixed effects, school fixed effects, and school trends, as well as the individual student covariates related to the cohort variables. All dependent variables are measured using Wave 3 of the Add Health. Test score is the standardized PVT score in Wave 3. "Ability grouping" is a dummy variable which is one if English classes in the cohort are grouped by ability or achievement. Wave 3 longitudinal weights are used. Figures in parentheses are standard errors robust to clustering at the school level. All controls are listed in Appendix Table 1.14 under individual characteristics, family characteristics, neighborhood characteristics and cohort controls. Variable definitions are in Appendix Table 1.12.



(a) Dependent variable: Registered to vote

(b) Dependent variable: Voted in 2000

Figure 1.1: Sensitivity of coefficients to measurement error in race variable

Notes: The y-axis variable is the average coefficient on the racial diversity index from 1,000 regressions where, before each regression, the race variable is replaced with a random value for a share of the sample. This share is indicated on the x-axis. This also affects the values of the racial diversity index and of the SES Gini index.

1.C. Tests for non-random clustering

Table 1.22: Tests for non-random clustering

	From Guryan et al. (2009)				
	White dummy	Black dummy	Hispanic dummy	Asian dummy	Other race dummy
Racial diversity index of peers in grade	−0.095	−0.004	0.045	0.026	0.027
	(0.087)	(0.047)	(0.044)	(0.036)	(0.032)
Racial diversity index of peers in school	172.496	−26.700	−52.317	−31.927	−61.551
	(22.341)	(11.773)	(10.432)	(5.175)	(6.998)
N	79,824	79,824	79,824	79,824	79,824
Adjusted R^2	0.505	0.360	0.326	0.184	0.127

	From Caeyers and Fafchamps (2020)				
	Transformed white dummy	Transformed black dummy	Transformed Hispanic dummy	Transformed Asian dummy	Transformed other race dummy
Racial diversity index of peers in grade	0.251	−0.059	−0.016	−0.033	−0.094
	(0.169)	(0.074)	(0.076)	(0.048)	(0.055)
N	79,824	79,824	79,824	79,824	79,824
Adjusted R^2	0.607	0.579	0.500	0.264	0.049

Notes: The table reports OLS estimates. Additional controls are school and grade fixed effects and school linear trends. The regressions reported in this table are run on the respondents to the Wave 1 in-school survey sample who are in cohorts containing at least one student in the estimation sample. Standard errors (in parentheses) are clustered at the school level. The data are unweighted.

Chapter 2

Betrayal aversion with and without a motive

Abstract

Previous studies find a preference for strategic risk over random risk for first movers in a trust game (Bohnet and Zeckhauser, 2004), but a preference for random risk over strategic risk in games of aligned interests. Using an experiment, we investigate whether removing the temptation payoff for the second mover in a trust game (and thus aligning players' interests) can overturn the strategic risk premium into a strategic risk discount.

We replicate the existence of a strategic risk premium in the trust game (known as *betrayal aversion*). We find no difference in preferences by type of risk when interests are aligned. We interpret these results as evidence that strategic risk premiums/discounts (including betrayal aversion) are reactions to the perceived intentions of the opponent. A second mover's intentions can be easily inferred from her actions in the trust game, but this is not the case in a game of aligned interests.

This chapter is co-authored with Martin Strobel.

I. Introduction

Several studies show that when making a risky decision people care not only about the objective probability distribution of outcomes, but also about whether the risk they face is random or strategic. Random risk is generated by a randomization device, while strategic risk is generated by a human opponent in a strategic interaction. How people value random versus strategic risk has been studied extensively in trust games (starting with the paper of Bohnet and Zeckhauser, 2004, henceforth BZ), where most studies find a strategic risk premium. That is, on average, people ask for a higher probability of success (reciprocated trust) to trust someone than to take an equiprobably risky bet. The strategic risk premium in this game has been dubbed "betrayal aversion".

In a paper which tries to understand what causes betrayal aversion, Bolton et al. (2016) find evidence that the degree of risk in a game—which is lower when players' interests are aligned—mediates how the nature of risk influences risk attitudes. Specifically, they find that in a game in which players' interests are aligned, there is the opposite of a strategic risk premium: a strategic risk discount.

In this paper, we examine whether aligning the players' interests in the trust game transforms the strategic risk premium into a strategic risk discount. In a between-subject design, we elicit risk attitudes towards strategic versus random risk in a trust game and in a similar but "toothless" game in which the second mover cannot gain an additional payoff by betraying the trust of the first mover. Unlike most previous papers, we inform participants that the strategic risk and the random risk are the same in a given game. We replicate the existence of a strategic risk premium in the trust game. However, in the game where players' interests are aligned, the type of risk—random or strategic—does not influence risk attitudes.

Our results indicate that betrayal aversion is a preemptive reaction to the opponent's perceived (malevolent) intentions, as suggested initially by BZ. Differences in valuing strategic versus random risk seem to exist only when the person who is the source of risk (the opponent) has a credible threat—that is, when she has a true alternative course of action. In our setting and after controlling for subjective beliefs about risk, aligning players' interests makes the

strategic risk premium vanish, but it does not transform it in a strategic risk discount.

This paper contributes to two strands of literature. The first one is the literature on betrayal aversion and the second one is the literature on the importance of intentions in strategic interactions.

As mentioned above, the literature on betrayal aversion started with BZ. The authors argue that differences in behavior when facing random versus strategic risk are potentially due to two factors: (1) the mere presence of another person (which may activate respondents' outcome-based or unconditional other-regarding preferences) and (2) that this person is responsible for the final outcome (such that intention-based or conditional other-regarding preferences could play a role).[1] To isolate the second factor, BZ compare first mover behavior in a binary trust game to that in an equiprobably risky social lottery.[2] Treatments differ in who makes the (equiprobable) decision at the second node: a person, or a randomization device. BZ demonstrate that, even after controlling for outcome-based social preferences, first movers require a substantial positive premium to enter a trusting relationship with a person, compared to playing an equally risky social lottery.[3] Most papers following suit replicate this result (Bohnet et al., 2008, 2010; Aimone and Houser, 2012; Aimone et al., 2015; Fairley et al., 2016; Quercia, 2016; Bacine and Eckel, 2018; Butler and Miller, 2018), but there is also evidence which questions the existence of this premium (Fetchenhauer and Dunning, 2012; Breuer and Hüwe, 2014).

BZ propose that the strategic risk premium (SRP) typically found for trusting decisions is due to an anticipated cost of betrayal, which they dub "betrayal aversion". The authors' preferred explanation is that betrayal aversion is a

[1] Models focusing on the outcome-based preferences approach are Levine (1998); Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Andreoni and Miller (2002); the simple model in Charness and Rabin (2002); Engelmann and Strobel (2004); Cox and Sadiraj (2007). There are two types of conditional preferences approaches: psychological game theory, which incorporates higher order beliefs into the utility function—examples of models are Rabin (1993); Dufwenberg and Kirchsteiger (2004)—and the "revealed intentions" approach, used by Cox et al. (2007, 2008). The model in Falk and Fischbacher (2006) and the general model in Charness and Rabin (2002) include both outcome-based preferences and intentions.

[2] A social lottery differs from a regular lottery by also having payoff consequences for someone who cannot influence the outcomes.

[3] In BZ, in the social lottery, first movers condition their entry on an average of 32% favorable outcomes, while in the trust game, they require on average that at least 54% of the second movers choose the favorable outcome.

preemptive action to shield first movers from the (malevolent) intentions of second movers. Several papers find support for this explanation. Across several treatments, Butler and Miller (2018) vary the degree to which second movers are aware of the consequences of their actions. They find that when second movers' actions do not reflect their intentions, the strategic risk premium turns into a strategic risk discount. Bolton et al. (2016) argue that if the perceived (malevolent) intentions of the second mover cause the SRP in the trust game, there should be no SRP in a game of aligned interests in which the only risk is miscoordination. The authors compare behavior in a stag hunt game and an equiprobably risky social lottery. In a between-subject design, they find a strategic risk discount. Bolton et al. (2016) propose that games act as frames, inducing either a trust or a distrust mindset: a setting with aligned interests makes first movers judge intentional actions of second movers more favorably. This explanation can reconcile both the SRP in the trust game and the strategic risk discount the authors find in the stag hunt game.

The original measure for betrayal aversion used by BZ (and also by Bolton et al., 2016; Butler and Miller, 2018) has been criticized for allowing for alternative explanations for the SRP (Li, 2020). Participants in the control game are not informed about how the risk they face is generated. This means that participants in the main games and those in the control games may imagine different distributions for these risks. This difference could affect behavior if participants violate the Substitution Axiom of expected utility theory. This is a common violation (Starmer, 2000). Li (2020) show that if this is the case, the strategic risk premium could reflect "ambiguity attitudes, complexity, different beliefs, and dynamic optimization". We measure betrayal aversion using a design adapted from Aimone and Houser (2012), which circumvents this issue (for details, see Section II).

Our paper also contributes to the literature on how motives and intentions are valued in strategic interactions. This literature suggests that eliminating the possibility to express unkind intentions does not necessarily induce positive reciprocity nor lead to a strategic risk discount. In our case, having a (weakly) Pareto-dominated option in the second mover's choice set in the game of aligned interests has no relevance for how kind the second mover's other

possible actions are perceived.[4] The idea that an action of the opponent is only informative about her intentions when there is a conflict between private and social interest is also captured in the model of Gul and Pesendorfer (2016).[5] On the experimental side, several papers find that positive reciprocity is stronger if a beneficial action by a first mover cannot be attributed to a strategic motive (Stanca et al., 2009; Johnsen and Kvaløy, 2016; Orhun; 2018; but see also Strassmair, 2009; Woods and Servátka, 2018). Ackfeld (2020) finds that people care not only about which action their opponent takes, but also about her reason for taking the action. The implication of this literature is that opponents' intentions will make less of a difference when the opponent does not have a credible threat than in one where she does. However, most of the experimental literature focuses on how first mover intentions are perceived by the second mover, while our paper's main focus is on how second mover intentions are perceived by a first mover who makes a decision conditional on second mover behavior.

Our paper bridges these two strands of literature by (i) using the same subject pool to study the willingness to take strategic versus random risks in two games which vary in the degree of complementarity between the two players' payoffs and by (ii) measuring strategic risk premiums (discounts) in a conservative way, to ensure they are not confounded by differences in beliefs about the degree of riskiness of a game.

The paper is structured as follows. Section II describes the experimental design. Section III explains the conceptual framework and presents the hypotheses. Section IV presents the data and the results. Section V discusses the results and their implications and concludes.

[4]Dufwenberg and Kirchsteiger (2019) refer to an option with extremely unfavorable payoffs for both players as a "murder/suicide" option, and Battigalli and Dufwenberg (2022) refer to it as a "bomb". As Battigalli and Dufwenberg (2022) put it, "while hurting everyone would surely be unkind, not doing so shouldn't automatically render other choices kind". The definitions of kindness proposed by Rabin (1993); Dufwenberg and Kirchsteiger (2004) and Dufwenberg and Kirchsteiger (2019) (a refinement of the definition in Dufwenberg and Kirchsteiger, 2004) ensure options are not perceived as kind when the alternative is a non-credible threat like "bomb".

[5]Gul and Pesendorfer (2016) relate their model to a large literature in experimental philosophy on what individuals classify as *intentional*. An action is classified as *non-intentional* if it benefits the agent while also being socially beneficial (Knobe, 2003).

II. Experimental design

Our workhorse is the version of the binary trust game used by BZ, shown in the upper right panel of Figure 2.1 (TG, for trust game).[6] The two players have equal endowments. The first mover (he) has two options. He can choose *Out*, which leaves both players with their initial endowment, or he can choose *In*. If he chooses *In*, he sends his endowment to the second mover. This endowment is doubled and the second mover (she) gets to choose between two options. These options are two ways of dividing the extra payoff between herself and the first mover. One option splits the amount equally between the two players (*Left*), while the other gives her a large share, leaving the first mover with less than his initial endowment (*Right*).

We modify this game along two dimensions: (i) whether or not the second mover has agency and (ii) whether or not option *Right* provides a higher payoff for the second mover than option *Left*. The second dimension means the two players play either a trust game or a game of aligned interests. This leads to a 2 x 2 design, with two games (a trust game and an aligned interests game) with two variants each (one involving strategic risk and the other involving random risk). Each participant is randomly assigned to one of the four treatments and to one role (first or second mover). The payoff structure for the first mover is identical in all treatments.

Figure 2.1 presents the treatments. We denote the four treatments TG (standard binary trust game), mTG (modified binary trust game)[7], AIG (aligned interests game) and mAIG (modified aligned interests game). TG and AIG involve strategic risk, mTG and mAIG involve random risk. We call second movers who have agency "active" (these are second movers in TG and AIG) and those who do not "passive" (second movers in mTG and mAIG).

Payoffs are expressed in lottery tickets. The first and the second payoff are for the first mover and for the second mover, respectively. The third payoff is the number of lottery tickets that are left unassigned: if this number is positive, there is a loss in efficiency. The bold underlined numbers emphasize that there is a positive temptation payoff for the second mover in the trust game, but not

[6]The first two payoffs are the the payoffs for the two players. We will explain the third payoff later in this section.

[7]This treatment is equivalent to the Risky Dictator game in BZ.

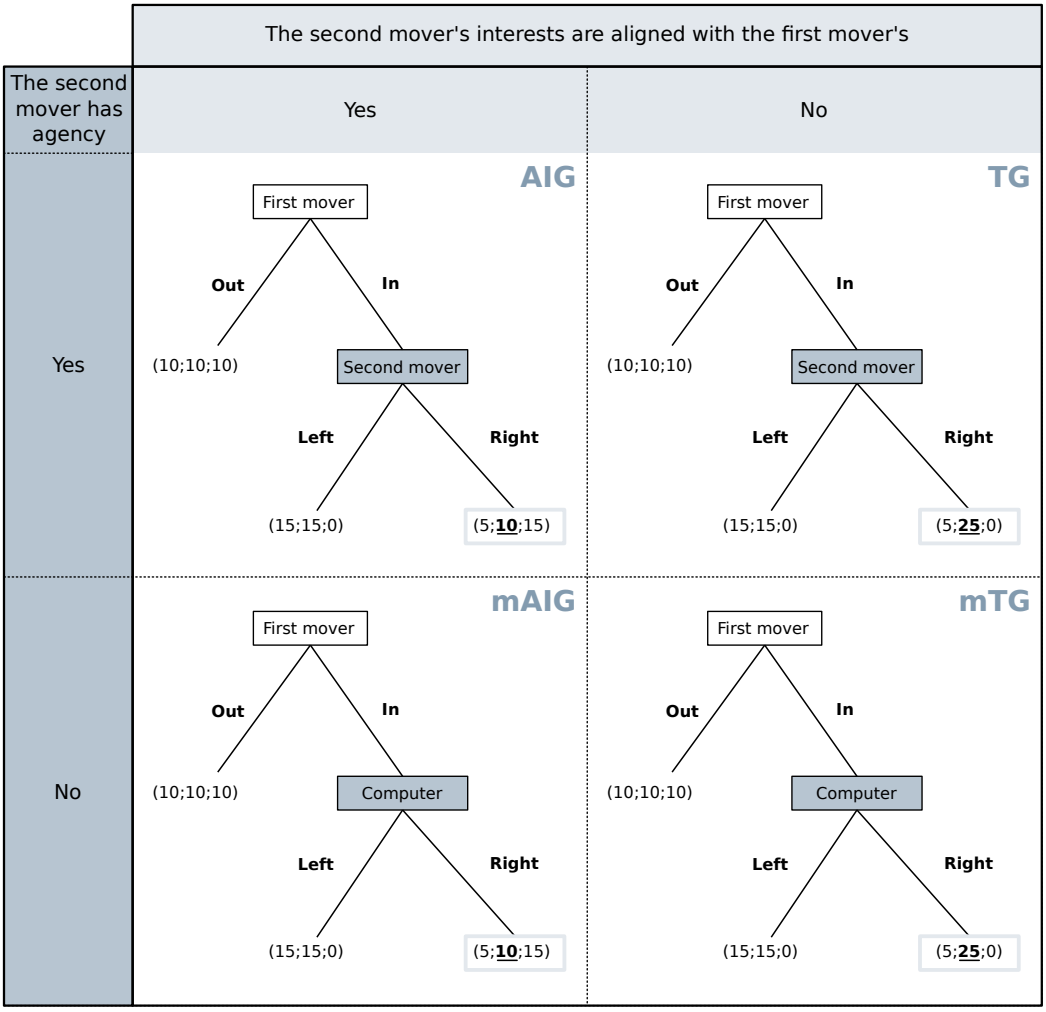


Figure 2.1: The treatments

in the aligned interests game. The temptation payoff is the extra payoff she can gain by choosing *Right* instead of *Left* when he chooses *In*.[8]

Participants make two decisions, one for a member of a different social group and one for a member of their own social group, in randomized order. A companion paper (corresponding to Chapter 3 in this dissertation) examines the role of group identity building on strategic risk preferences. In this paper, we

[8]We wanted to ensure that participants understand that the outcomes differ in efficiency. Since payoffs are in lottery tickets, they might focus on their payoff relative to the opponent's when ordering outcomes. For instance, participants might wrongly conclude that *In, Right* is better for the second mover than *In, Left* in the game of aligned interests: she has double his chances to win in the former (she receives 10 tickets, he—5), but they have equal chances to win in the latter (she receives 15, he—15 as well). Adding unassigned tickets solves this issue by aligning the ordering of relative chances to win to the ordering of absolute chances.

pool all decisions and abstract from the identity of the opponent. This does not pose a problem for our results, as decisions reported in this experiment are made before group identity had an impact on behavior.[9]

We collected decisions simultaneously using the strategy method (Selten, 1967). Active second movers in TG and AIG chose between *Left* and *Right*, in case their assigned first mover chose *In*. Passive second movers in mTG and mAIG did not make any decision. First movers in TG and AIG indicated the minimum percentage of second movers in their treatment who would have to choose *Left* for them to prefer *In* over *Out*. In mTG and mAIG, the computer moved randomly, according to the same distribution as in TG and AIG, respectively (p^*, *Left*; $1 - p^*$, *Right*). First movers stated the minimum percentage of *Left* options in those distributions to prefer *In* over *Out*. First movers' choice is called the *minimum acceptable probability* (MAP). It is a percentage between 0 and 100 (non-integers were allowed).

We ran all treatments in parallel. Afterwards we computed the percentage p^* of *Left* decisions made by second movers in the relevant pool for each first mover. If his MAP was higher than p^*, we implemented *Out*. Otherwise we implemented *In*. In the latter case, the decision of a randomly chosen relevant second mover determined the payoffs.

We were interested in first mover decisions. The MAP represents their indifference point between the sure option *Out* and entering a lottery with externalities for another player (in mTG/mAIG)—or an equiprobable game with another player (in TG/AIG). The SRP in each game is the difference between the mean MAP in the variant played against an active second mover (TG or AIG) and the mean MAP in the variant played against a passive second mover (mTG or mAIG).

We varied second movers' agency in the trust game to replicate BZ. The same variation in the aligned interests game allows us to study how first movers' response changes when the opportunity set of the second movers no longer includes an action with a temptation payoff (or a motive to betray). This shows how the SRP changes with the possibility to attribute various intentions to the

[9] In the regressions reported in this paper using the pooled sample for all four treatments, we do not observe any differences in behavior towards in- versus outgroup members, nor between first versus second decisions. However, if we run mean-comparison tests on more specialized pools (only first decisions, only for ingroup opponents), we do find some differences.

second mover based on the game structure.

II.A. Protocol

The experiment was conducted in the BEELab experimental laboratory at Maastricht University using Qualtrics in September–October 2017. Participants were first year business and economics students who were recruited via a compulsory course. We selected this subject pool to meet the requirements for the companion paper's matching protocol (for details, see Chapter 3).

First movers and active second movers made two decisions, one for an in- and one for an outgroup member, with no feedback in between. Subjects knew that one of these two decisions would be randomly drawn to determine their payoff.[10]

Table 2.1: Participants by treatment

Game	Treatment	Role	# assigned subjects	# subjects with minor/no understanding mistakes	% subjects with minor/no understanding mistakes
Trust game	TG	First mover	92	41	45%
		Second mover	34	20	59%
	mTG	First mover	81	24	30%
		Second mover	27	-	-
Aligned interests game	AIG	First mover	78	23	29%
		Second mover	35	23	66%
	mAIG	First mover	79	23	29%
		Second mover	19	-	-

Table 2.1 shows the number of participants in each treatment-role combination. There were 445 participants (330 assigned to first mover roles) across 69 sessions, with a varying number of subjects per session (1–13).[11,12]

Passive second movers were briefly informed about the procedure and were allowed to leave without making any decision. Active second movers and first

[10]The matching procedure is explained in detail in Appendix 2.A.

[11]This is because subjects also participated in one of two other experiments which differed in capacity constraints. The two experiments were about consumer behavior and product choice and involved no social interaction between subjects.

[12]296 additional students who were also taking the course in which we recruited our subjects participated as passive second movers (not shown in Table 2.1). They did not have to make any decision and were not present in the sessions conducted for this experiment. In total, 741 students took part in the lottery (details follow later in this section).

movers had to first answer a couple of comprehension questions. After that, they had to make two decisions corresponding to the treatment-role combination to which they had been assigned, one for an out- and one for an ingroup member. In a third part, we elicited risk preferences, generalized trust, and positive and negative reciprocity using survey preference measures developed by Falk et al. (2016), basic demographics, and a measure of attachment to their social group.

The experiment took on average 20 minutes. This was a time limit imposed by the respondents' participation in one of two other unrelated experiments. Because of this time constraint, participants were allowed to go through the comprehension questions only once. They were offered immediate feedback with the correct answers and were allowed to continue regardless of whether they had answered correctly or not. Participants could ask questions privately at any time during the experiment. Since we cannot guarantee that everyone understood the instructions correctly, we take a conservative approach and restrict the analysis to the sample of first movers who answered the understanding questions correctly or made a minor mistake (Table 2.1 presents the sizes of the full samples and of the estimation samples).[13,14]

[13]We accepted the following minor mistakes: (1) when asked what their payoff would be in a certain situation, some participants did not report the correct payoff, but the sum of the correct payoff and the show-up fee; (2) in games against passive opponents, first movers were asked the value of p^* if 35% of active second movers in the corresponding game chose *Right*. We considered that those who answered 35% (instead of 65%) might have not paid attention, so we attributed this to a minor mistake.

Dropping criterion (2) decreases the estimation sample by 8 participants. If we drop these 8 participants from the estimation sample, the coefficients distinguishing between game variants in Table 2.3 keep their sign, but are not significant at conventional levels.

[14]This low rate of subjects' understanding of the instructions from the first try (111 in 330) is not uncommon in betrayal aversion experiments, which are rather complex. For instance, in a paper using the same elicitation method and similar comprehension questions, Quercia (2016) mentions similar rates of subjects who answer the questions correctly (see table 3, column 2 on p. 57).

A potential concern is that the cross-treatment comparisons we report could be affected by selection bias if participants are systematically more likely to answer the understanding questions correctly in one treatment than in another. In Appendix Table 2.4 we check whether those who answer the comprehension questions correctly are similar in terms of observable characteristics to those who do not (in terms of risk aversion, sex, but also assigned treatment: TG versus mTG, AIG versus mAIG, TG versus AIG, mTG versus mAIG). We find no significant differences within games. However, when comparing those who face an active opponent in TG versus AIG, those in TG are more likely to answer correctly than those in AIG if we do not include session fixed effects and do not adjust for multiple comparisons. While this suggests that comparisons between behavior in TG and AIG might be affected by selection bias, none of the coefficients

Subjects received course credit for participation. On top of that, they received a show-up fee of 3 lottery tickets and the lottery tickets they earned in the experiment. The lottery tickets offered the chance to win one of 15 widely used vouchers worth €100. On average, each participant had a 1.75% chance to win a voucher, making the average expected payoff €1.75. The average expected payoff was higher for first movers and for active second movers (who were the ones spending 20 minutes in the laboratory) at €2.33 (approximately €7/hour).[15,16]

We present experimental instructions in Appendix 2.B.

III. Hypotheses

Let MAP_{TG}, MAP_{mTG}, MAP_{AIG}, and MAP_{mAIG} be the minimum acceptable probabilities in the four treatments. For each game—the trust game or the aligned interests game—we define the strategic risk premium in that game as

$$SRP_{game} = MAP_{active\ opponent} - MAP_{passive\ opponent}$$

Similarly to the majority of papers on betrayal aversion mentioned in Section I, we expect to find a positive SRP in the trust game.

Hypothesis 1 $SRP_{TG} > 0$.

Based on previous findings that in games of aligned interests players prefer strategic to random risk (Bolton et al., 2016), we expect to find a strategic risk discount in the game of aligned interests.

Hypothesis 2 $SRP_{AIG} < 0$.

We also state a weaker hypothesis than hypotheses 1 and 2 taken together. Based on the literature on the valuation of intentions mentioned in Section I, we assume the following about the magnitude of the first mover's reaction to

remains significant if we adjust for multiple comparisons using a Yekutieli-Benjamini correction (Yekutieli and Benjamini, 2001).

[15]Actual expected earnings were computed after data collection and matching. In the instructions, participants were told that approximately 700 students would take part in the experiment and that they could win one of the 15 vouchers worth €100 each.

[16]This amount is slightly higher than the *net* minimum hourly wage for a 20-year old in the Netherlands at the time. The *gross* minimum hourly wage for a 20-year old in 2017 was between €6.33 and €7.03, depending on the sector. Source: `https://rijksoverheid.sitearchief.nl/#archive`. Type "minimumloon 2017" in the search box (in Dutch; accessed April 26, 2022).

the perceived intentions of the second mover: a choice of *Right* by the second mover in the trust game unequivocally expresses malevolent intentions: she hurts the payoff of the first mover and profits from it. A choice of *Left* in the same game shows benevolent intentions: she forgoes a higher payoff for his benefit, following his move to go for efficiency (*In*) at the cost of risking his payoff. In the aligned interests game, the second mover's intentions are not as clear, as *Left* benefits both players, while *Right* hurts both. If what distinguishes settings with random risk from those with strategic risk are the intentions of the opponent, the SRP should be closer to zero when the opponent's intentions given an action are unclear (such as in the aligned interests game) than when they are clear (such as in the trust game).

Hypothesis 3 $|SRP_{TG}| > |SRP_{AIG}|$.

Depending on the signs of SRP_{TG} and SRP_{AIG}, we will be able to specify Hypothesis 3 more precisely.

IV. Data and results

Table 2.2 displays descriptive statistics of minimum acceptable probabilities of first movers in the four treatments. P-values in the table are from two-sided Mann-Whitney tests.

We first look at comparisons between sources of risk within a game. At first glance, data in Table 2.2 seem to weakly support Hypothesis 1. Mean MAPs in all three panels are higher against active than against passive opponents in the trust game. The comparisons in Table 2.2 do not take into account that the two decisions made by each first mover might be correlated.

To address this, we compute the mean MAP for each subject. The mean MAP is not significantly different in TG versus in mTG (p-value = 0.22, two-sided Mann-Whitney test), nor in AIG versus mAIG (p-value = 0.43, two-sided Mann-Whitney test).[17]

We also run an OLS regression analysis in which we account for the potential correlation between the two MAP decisions made by each subject by clustering standard errors at the individual level. Regression results are in Ta-

[17] If we only compare MAPs stated as first decision in TG versus mTG, the difference is significant (p-value = 0.05, two-sided Mann-Whitney test), but remains insignificant in AIG versus mAIG (p-value = 0.69, two-sided Mann-Whitney test) (see the middle panel in Table 2.2).

Table 2.2: Minimum Acceptable Probabilities across treatments

	Both decisions, pooled			
	TG	mTG	AIG	mAIG
MAP	61.15	52.67	60.65	63.30
	(18.99)	(24.03)	(19.18)	(16.28)
p-values	0.071		0.359	
Observations	82	48	46	46

	First decision			
	TG	mTG	AIG	mAIG
MAP	62.44	51.00	60.65	60.91
	(15.99)	(22.79)	(18.05)	(15.52)
p-values	0.049		0.691	
Observations	41	24	23	23

	Second decision			
	TG	mTG	AIG	mAIG
MAP	59.87	54.33	60.65	65.70
	(21.72)	(25.58)	(20.65)	(17.01)
p-values	0.490		0.343	
Observations	41	24	23	23
Individuals	41	24	23	23

Notes: The table shows averages per game variant, either with an active or with a passive opponent. Each participant made two decisions. p-values are from ranksum tests of opponent type (active or passive) within a game. Standard deviations in parentheses.

ble 2.3. Models (1) and (2) use the full sample of first movers with minor or no understanding mistakes, without and with individual controls, respectively. Because in our estimation sample we only kept respondents who showed an understanding of the instructions, and since in a couple of sessions there were few participants, in 15 cases there is only one participant in the estimation sample per experimental session. This prevents us from including session fixed effects with this sample; models (3) and (4) drop these 15 individuals to allow for session fixed effects. Model (3) includes session fixed effects, but no individual controls; model (4) includes both individual controls and session fixed effects.

To test Hypothesis 1, we have to look at the coefficient of *Active second mover* in Table 2.3. This coefficient quantifies the SRP in the trust game, since the baseline is the MAP against a passive second mover in the trust game (MAP_{mTG}). Models (1) and (2) show a SRP close to being marginally significant (p-value = 0.10 in model (1) and p-value = 0.13 in model (2)); the models with session fixed effects indicate more strongly that there is a positive SRP, with p-value = 0.03 in model (3) and a p-value = 0.04 in model (4). We thus find evidence supporting Hypothesis 1.

Result 1 We find weak evidence of a positive SRP in the trust game.

To test Hypothesis 2, we test whether the linear combination *Active SM* $\times$ *Aligned interests game* differs significantly from zero. This is not the case (for instance, p-value = 0.85 in model (4)). We cannot conclude that there is either a strategic risk premium, nor a discount in the aligned interests game.

Result 2 We do not find evidence of a strategic risk discount in the aligned interests game.

Table 2.3: Linear regressions on Minimum Acceptable Probabilities

	(1)	(2)	(3)	(4)
Baseline: mTG				
Active second mover (**H1**)	8.49	6.92	11.12* *	10.89* *
	(5.17)	(5.11)	(5.03)	(5.44)
Aligned interests game	10.64*	8.67	11.76* *	10.61*
	(5.39)	(5.52)	(5.84)	(6.02)
Active second mover × Aligned interests game	−11.14	−8.51	−13.64*	−11.95
	(6.89)	(7.02)	(7.29)	(7.91)
Risk aversion (0–10)		−1.77*		−1.57
		(0.94)		(1.11)
Male		2.79		0.40
		(3.47)		(4.19)
Ingroup first		0.46		−1.77
		(3.38)		(3.88)
Constant	52.67* **	65.23* **	46.06* **	59.85* **
	(4.53)	(7.55)	(5.20)	(10.31)
Linear & nonlinear combinations				
Active SM + Active SM × Aligned interests game (**H2**)	−2.65	−1.60	−2.51	−1.06
	(4.56)	(4.76)	(5.24)	(5.65)
|Active SM| – |Active SM + Active SM × Aligned interests game| (**H3**)	5.84	5.32	8.61	9.83
	(6.89)	(6.95)	(7.23)	(7.77)
Aligned interests game + Active SM × Aligned interests game	−0.50	0.15	−1.88	−1.34
	(4.29)	(4.28)	(4.33)	(4.71)
Session fixed effects	No	No	Yes	Yes
Adjusted R^2	0.02	0.03	0.11	0.11
Observations	222	222	192	192
Individuals	111	111	96	96
Sessions	41	41	26	26

Notes: Standard errors clustered at individual level in parentheses. 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. This is why, 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

To test Hypothesis 3, we examine the coefficient of a nonlinear combination, |Active SM| - |Active SM + Active SM $\times$ Aligned interests game|. This coefficient does not significantly differ from zero in any specification (p-value = 0.21 in model (4)).

Result 3 We do not find evidence that strategic risk premiums in the two games differ significantly from each other in absolute value.

As an exploratory analysis, we also compare behavior between games. While we do not have *a priori* hypotheses about this, results are informative about the effects of an institution which would eliminate the temptation payoff for the second mover. The p-value for the two-sided Mann-Whitney test comparing first mover behavior in TG and AIG is 0.94. When looking at the mean MAP per subject, MAP_{mTG} is marginally lower than MAP_{mAIG} (p-value = 0.06, Mann-Whitney test). For games against active second movers, the difference between the mean MAP_{TG} versus MAP_{AIG} is not significant (p-value = 0.99, Mann-Whitney test).

For games against active second movers, we look at the coefficient of *Aligned interests game + Active SM $\times$ Aligned interests game*. It is not significant in any of the four specifications (in (4), p-value = 0.78). For games against passive second movers, we examine the coefficient of *Aligned interests game*. This is marginally significant in (1) and (4) and significant in (3) (p-value = 0.05).[18]

Result 4 There is a marginally higher MAP in mTG compared to mAIG.

Result 5 We do not find a significant difference between the MAPs in TG versus AIG.

While not the focus of this study, we take a brief look at second mover behavior. By moving from the trust game to the aligned interests game, p^* (the share of second movers who choose the cooperative action *Left*) increased from 60% to 87% among second movers with minor or no understanding mistakes. This simply shows that second movers reacted to the different incentives in the two games. Even if the MAPs in TG and AIG are similar, the effect of eliminating the temptation payoff on the probability of cooperation of second movers

[18]Except for the (non)existence of the temptation payoff, the trust game and the aligned interests game also differ in terms of efficiency. In the game of aligned interests, the second mover choosing *Right* leads to 15 lottery tickets being wasted. While this does not affect within-game comparisons (TG versus mTG, AIG versus mAIG), it is possible that some of the across-game differences (TG versus AIG, mTG versus mAIG) are due to efficiency concerns.

was so dramatic that it substantially increased realized strategic interactions. It did so however not through reducing the standards required by first movers to enter strategic interaction (their MAPs), but exclusively by making second movers choose *Left* more often. We conclude that in order to harness reactions to second movers' intentions on the part of first movers (and thus affect their MAP), these intentions have to be easily deducible in the context of the game. In AIG, since the choice of *Left* is the dominant strategy for second movers, it is uninformative about the chance a first mover faces a kind/unkind opponent.

V. Discussion and conclusion

In this paper, we find that eliminating the motive of the second mover to betray the first mover in a binary trust game does not lead to a SRP reversal, as we had hypothesized based on the results of Bolton et al. (2016) and Butler and Miller (2018), who find evidence in line with such a reversal. Our finding supports the idea that it is not sufficient to align players' interests to observe a strategic risk discount. When first movers cannot properly identify their opponent's type (kind or unkind), they are indifferent between having a human make a decision that influences their payoff and taking an equiprobable gamble with payoff externalities for another player. We interpret this as evidence that intentions matter when they can be read, that is, in contexts without "muddied motivations".

There are several differences between the experimental design used in this paper and the ones in Bolton et al. (2016) and Butler and Miller (2018). A plausible reason for the differences in findings between this paper and those in Bolton et al. (2016) and Butler and Miller (2018) is that the probabilities p^* of the favorable outcome for the first mover were communicated differently. As Li (2020, Section 2 and Appendix A) note, in BZ and in subsequent studies that aimed at eliminating ambiguity (such as Bolton et al., 2016; Butler and Miller, 2018), ambiguity in the sense of modern ambiguity theories might still have been present and is a potential confounding factor. These experiments presented the way the probability p^* of a favorable outcome was generated differently to subjects facing active versus passive second movers. For subjects with active opponents, p^* was determined from the unknown distribution of

active second movers' choices.[19] For those with passive opponents, p^* was either drawn from an explicitly uniform distribution (as in Bolton et al., 2016, see p. 418), or from a distribution about which first movers did not know how it had been generated (as in Bohnet and Zeckhauser, 2004; Butler and Miller, 2018). It is thus plausible that in the second case first movers had no reason to expect a non-uniform distribution, and highly unlikely that they expected exactly the same distribution that they would have expected had they played against an active second mover. This allows for the possibility that, should first movers not be rational expected utility maximizers, differences in first mover behavior between treatments in these papers come from a change in the expected p^* (or its distribution).

The experimental design used in this study, adapted from Aimone and Houser (2012), overcomes this issue by explaining to subjects facing both types of second movers that the probability of the favorable outcome p^* was determined in the same way for both treatments in a game, using the (unknown) distribution of active second mover decisions. In other words, while ambiguity might still be present, observed behavioral differences between treatments *in the same game* cannot be attributed to differences in ambiguity.[20]

We now turn to the between-games comparison. We have contrasted an institution in which betrayal is possible with a secure institution that removes the temptation payoff from betrayal for the second mover. We were interested in whether the conditional strategic risk premium would be lower in absolute value in the game of aligned interests compared to the trust game. Should this have been the case, then removing the temptation payoff would have increased the percent of time players interacted with each other (i.e. the percent of time *In* was played out instead of *Out*). This increase could happen through two channels: (1) by increasing the percentage of second movers who choose the efficiency maximizing action *Left* in the game of aligned interests relative to the trust game and (2) by lowering the conditional threshold set by first movers to prefer *In* to *Out* in the game of aligned interests relative to the trust game. Previous literature suggested that (2) could arise as a by-product of

[19] Technically speaking, since first movers had already been matched when making their decisions, this refers to the probability that *their already matched opponent* would choose *Left*.

[20] This holds under the assumption that ambiguity aversion is constant within a game, whether it is played against an active or against a passive opponent.

the second mover's intentions being viewed favorably in the game of aligned interests. However, as the literature on the valuation of intentions illustrates, a policy aimed at increasing welfare by eliminating the temptation payoff for second movers might not result in (2), as first movers could adapt the conditions under which they are willing to enter the new transactions. In this experiment, eliminating the motive of the second mover to betray does not increase first movers' willingness to enter in games against active second movers ($MAP_{TG} \approx MAP_{AIG}$), while it even marginally decreases it in games against passive movers ($MAP_{mTG} < MAP_{mAIG}$). This makes the strategic risk premium in the trust game not significantly different from the one in the aligned interests game. Because p^* is close to one in mAIG/AIG (nearly all second movers in AIG choose the welfare maximizing option *Left*), the number of realized strategic interactions is anyhow higher in the aligned interests game. This increase in the number of realized transactions is due to vastly more second movers behaving cooperatively (channel (1)), and thus more of them living up to the (same) standards imposed by first movers.

To sum up, we find a strategic risk premium in the trust game. We also find no influence of the type of risk (random or strategic) on the willingness to take risks in a game of aligned interests. More subjects end up taking the riskier route with an active human opponent when the riskiness of this option is lower. This is not due to a higher willingness to take risks in that setting, but to more opponents responding to the change in incentives and chosing the cooperative option more often.

In light of these findings, we believe that an interesting further step would be to examine how betrayal aversion varies in response to more fine-grained changes in the temptation payoff, and whether players' "betrayal sensitivity" is a preference that carries over to other contexts. Is betrayal aversion a feature of the situation or a characteristic of a person?[21] Another direction is to try to identify the necessary *and* sufficient conditions for strategic risk discounts to come about. A third direction is to link the two directions above with research on betrayal under ambiguity. Such findings would be more readily applicable to the vast majority of settings outside the laboratory (Li, 2020).

[21]Aimone et al. (2015) measure intra-individual betrayal aversion and find no correlation with risk aversion. To identify a player's betrayal sensitivity, one needs several measurements per individual, in games with several temptation payoffs to the second mover.

From a practical perspective, our findings imply that institutional changes such as removing the temptation payoff are limited in scope. While we find that second movers respond to the incentive change and act more cooperatively, we fail to find the additional effect that first movers are more responsive to such a change when facing strategic as opposed to random risk. Different perceptions of ambiguity in the strategic versus the random risk setting in earlier literature could explain findings suggesting otherwise.

Appendix

2.A. Matching procedure

Data collected for this study partially overlaps with data used in a companion paper (corresponding to Chapter 3 in this dissertation). The companion paper studies the impact of social identity on betrayal aversion. For this reason, the matching procedure for the entire dataset 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.

Members of the social groups were spread unevenly across experimental sessions. We assigned the first ten individuals in a social group in show-up order to second mover roles and from then onward, in round-robin fashion within each social group to first mover roles. Out of the first ten participants in a social group, roughly the first four were assigned passive second mover roles and the next six were assigned active second mover roles.[22] As a consequence, the actual matching was not done during the experimental sessions, but after all data had been collected, according to a matching rule decided upon in advance.

The matching procedure ensured that:

– first movers knew that their randomly drawn decision may also affect the payoff of *another participant*;

– active second movers knew that their decision may affect the payoff of *other participants*;[23]

– passive second movers knew their payoff was determined either by a com-

[22]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. We control for decision order in our analysis; the paper in Chapter 3 discusses this aspect in detail.

[23]We chose for this asymmetry in the instructions for first movers and active second movers because we were interested solely in first mover decisions. We 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.

puter draw, or jointly by a computer draw and another participant's decision.

Participants received sheets with unique randomly generated four-digit codes, which they had to present in order to collect their earnings a couple of weeks after the experiment from a third party not involved in running the experiment. Within each treatment-role-match type pool of subjects (in-/outgroup), 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 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.[24]

[24]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.

2.B. Instructions

2.B.A. Instructions on paper

Participants were randomly assigned to cubicles in the laboratory. Then, they had to type their randomly assigned participation number and their community identifier ("community" was the name under which their social group was known). They were assigned to a role depending on this community identifier, as explained in Appendix 2.A. After this, they were handed the paper instructions corresponding to their role. They were instructed to read the paper instructions before clicking "Start" to begin the experiment.

We present the instructions for the roles in the trust game and in the modified trust game. The corresponding instructions for the aligned interests game and the modified aligned interests game differ only with respect to the payoffs in the decision tree.

□ ***First movers in the [modified] trust game***

Thank you for participating in this study. We will conduct a lottery among all participants (roughly 700). You can win one of 15 shopping vouchers worth 100 Euros each. These vouchers are accepted by many Dutch shops, including online stores. You receive 3 lottery tickets for your participation. During the study, you may earn more tickets depending on your decisions. The more lottery tickets you have, the higher the chance to win a voucher.

The draw will take place at the end of block period 1. Information about the winning lottery codes as well as feedback about the study will be published on *[the learning platform]*. The last page of these instructions contains your **lottery code**: please detach it and take it with you when you leave. If you win you have to show this page in order to receive your prize. If you lose it, we can't hand you the prize.

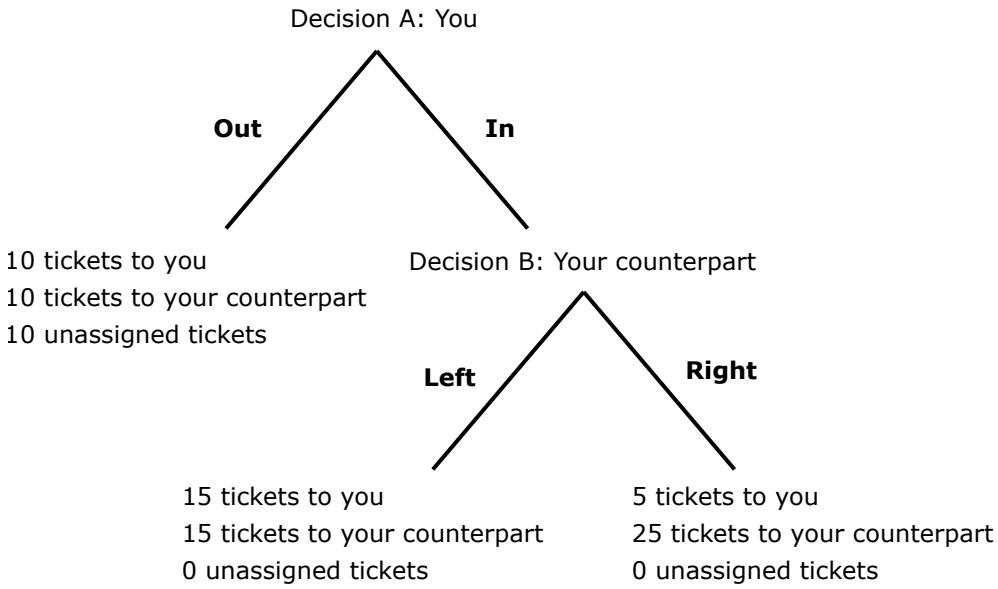
You will have to answer two questions that may affect your earnings and those of another participant. After all participants have answered, one of your two answers will be randomly drawn and chosen to determine your earnings. Both answers are potentially important, because you don't yet know which one will be implemented. The two questions are about situations which are similar, but not identical, so please read the instructions carefully. All earnings in this study are expressed in lottery tickets.

Your answers in this study are strictly confidential. Your unique lottery code is your identifier: without knowing it, no one can trace your answers. There is no communication among participants. After reading the instructions, click **Start** on the computer screen to start the study. The questions will appear on the screen.

The decision situation

We first introduce you to the basic decision situation. You are randomly matched with another participant: your counterpart. You and your counterpart can earn a maximum of 30 lottery tickets together, depending on your decisions. You have to take Decision A, in which you are confronted with two alternatives, **In** or **Out**. If you choose **Out**, you and your counterpart get 10 lottery tickets each and 10 lottery tickets are left unassigned. If an unassigned ticket is drawn, the corresponding prize is not awarded to anyone. If you choose **In**, the outcome depends on *your counterpart's Decision B [the random Decision B taken by a computer]*.

Your counterpart [The computer] chooses between **Left** and **Right**. If *he/she [the computer]* chooses **Left**, you and your counterpart get 15 lottery tickets each and there are no unassigned tickets. If instead, *he/she [the computer]* chooses **Right**, you receive 5 lottery tickets, your counterpart receives 25 lottery tickets and there are no unassigned tickets.



[Only for the modified trust game:]

The computer decides by randomly picking a card from a deck of cards; the cards in the deck are marked either "Left" or "Right". The computer's decision coincides with what is marked on the randomly drawn card. The deck of cards has been compiled in the following way. In a game similar to the one just described, the same Decision A and Decision B as above were made by two randomly matched human

participants (instead of a human and a computer, respectively). The decisions in this other game had the same payoff consequences for the two participants as they have for you and your counterpart. Before knowing what their match had chosen for Decision A, participants who had to make Decision B had to answer the following question: "Which option, Left or Right, do you choose in case your counterpart chooses In?" For each Decision B to choose Left in this other game, a Left card has been added to the deck in your game. For each Decision B to choose Right, a Right card has been added to the deck.

The study

The study is based on the decision situation just described to you.

[Only for the trust game:]

In this study, participants can take one of two decisions: Decision A, like you, or Decision B, like your counterpart. We will ask you to take your decision considering not only the possible action of your counterpart but also the possible actions of all participants taking Decision B. In particular, we will ask you to answer the following question:

KEY QUESTION: How large would the percentage p of *participants taking decision B who chose [cards marked]* Left minimally have to be for you to pick In over Out? (the number must lie between 0 and 100)

[Only for the trust game:]

Before knowing what their match had chosen for Decision A, participants who had to take Decision B had to answer the following question: "Which option, Left or Right, do you choose in case your counterpart chooses In?"

After all participants have made their choices, we will calculate the percentage of *participants taking Decision B who chose Left [cards marked Left in the deck compiled by the computer]*, let's call it p^*. If p^* is greater than or equal to your required value of p (from your answer to the KEY QUESTION above), your earnings for this decision will be determined by *your counterpart's [the computer's]* Decision B. If p^* is less than your required value of p (your answer to the KEY QUESTION above), you and your counterpart will get 10 lottery tickets each and 10 tickets will be unassigned.

Two examples should make this clear. Note: the numbers used below are only examples and are not necessarily representative of the *Decisions B taken by participants [composition of the deck of cards]*.

EXAMPLE 1: Suppose 95% of Decisions B are Left, that is, p^* is 95%. Suppose further that your answer to the KEY QUESTION, p, is 40%. Since p^* is greater than p, your Decision A would be **In**. At that point, there would be two possible cases: either *your counterpart [the card drawn by the computer]* is among the 95% of *those who chose [the cards marked]* **Left** or he/she is among the 5% *who chose [the cards marked]* **Right**. In the former case, you and your counterpart would get 15 lottery tickets each and there would be no unassigned tickets. In the latter case, you would get 5 lottery tickets, your counterpart would get 25

lottery tickets and there would be no unassigned tickets.
EXAMPLE 2: Suppose 5% of Decisions B are **Left**, that is, p^* is 5% and suppose further that your answer to the KEY QUESTION, p, is 60%. Since p^* is lower than p, your Decision A would be **Out**. In this case, you and your counterpart would get 10 lottery tickets each and 10 tickets would be unassigned, regardless of whether *your counterpart's [the computer's]* choice is **Left** or **Right**.
Before making any decision we ask you to complete a quiz to check your understanding of the instructions. To start the quiz, please click **Start** on the computer screen.

□ ***Second movers in the trust game***

Thank you for participating in this study. We will conduct a lottery among all participants (roughly 700). You can win one of 15 shopping vouchers worth 100 Euros each. These vouchers are accepted by many Dutch shops, including online stores. You receive 3 lottery tickets for your participation. During the study, you may earn more tickets depending on your decisions. The more lottery tickets you have, the higher the chance to win a voucher.

The draw will take place at the end of block period 1. Information about the winning lottery codes as well as feedback for the study will be published on *[the learning platform]*. The last page of these instructions contains your **lottery code**: please detach it and take it with you when you leave. If you win you have to show this page in order to receive your prize. If you lose it, we can't hand you the prize.

You will have to answer two questions that may affect your earnings and those of other participants. After all participants have answered, one of your two answers will be randomly drawn and chosen to determine your earnings. Both answers are potentially important, because you don't yet know which one will be implemented. The two questions are about situations which are similar, but not identical, so please read the instructions carefully. All earnings in this study are expressed in lottery tickets.

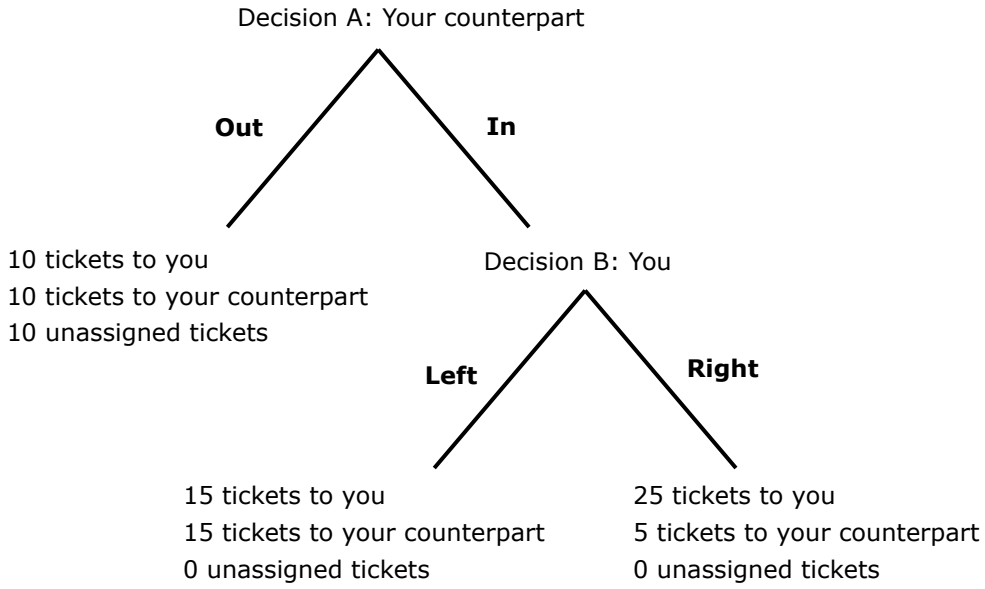
Your answers in this study are strictly confidential. Your unique lottery code is your identifier: without knowing it, no one can trace your answers. There is no communication among participants.

After reading the instructions, click **Start** on the computer screen to start the study. The questions will appear on the screen.

The decision situation

We first introduce you to the basic decision situation. You are randomly matched with another participant: your counterpart. You and your counterpart can earn a maximum of 30 lottery tickets together, depending on your decisions. Your counterpart has to take Decision A, in which he/she is confronted with two alternatives, **In** or **Out**. If he/she chooses **Out**, you and your counterpart get 10 lottery tickets each and 10 lottery tickets are left unassigned. If an unassigned ticket is drawn, the corresponding prize is not awarded to anyone. If he/she chooses **In**, the outcome depends on your Decision B.

You choose between **Left** and **Right**. If you choose **Left**, you and your counterpart get 15 lottery tickets each and there are no unassigned tickets. If instead, you choose **Right**, you receive 25 lottery tickets, your counterpart receives 5 lottery tickets and there are no unassigned tickets.



In this study, we will ask you to answer the following question:

KEY QUESTION: Which option, Left or Right, do you choose in case your counterpart chooses In?

Before making any decision we ask you to complete a quiz to check your understanding of the instructions. To start the quiz, please click **Start** on the computer screen.

□ ***Second movers in the modified trust game***

Thank you for participating in this experiment series. We will conduct a lottery among all participants (roughly 700). You can win one of 15 shopping vouchers worth 100 Euros each. These vouchers are accepted by many Dutch shops, including online stores.

Your chance of winning has been determined by the choice of a computer alone or by the combined choices of another participant and of a computer. The draw will take place at the end of block period 1 and the 15 winning lottery codes will be published on *[the learning platform]*.

2.B.B. Instructions on screen

□ ***All first movers and second movers in the trust game***

This study has three parts: the quiz, the key questions and a short survey. Each part will be marked clearly on the screen.

Quiz

The next part contains several quiz questions to confirm you perfectly understand the instructions. Your answers to these questions do not influence anyone's earnings. The software will check whether your answers are correct and inform you about it. If your answers are correct, you may continue to the following part. If there are any mistakes, please place the white sheet on the door of the cubicle and an instructor will come to explain.

Key questions

After you have successfully answered the quiz questions, you will have to answer two clearly marked KEY QUESTIONs. YOUR ANSWERs to these questions will determine your earnings and may determine *[first movers:] another participant's [second movers: other participants']* earnings.

Survey

The last part of the study is a short survey.

If you have questions at any point, please place the white sheet on the door of the cubicle and an instructor will come to respond.

Please read the paper instructions fully. Once you are done, you may start the quiz.

□ ***First movers in the [modified] trust game***

Quiz

1) Assume you stated that the minimum p for you to choose **In** over **Out** is 20% and p* is 30%. Further assume *your counterpart chose [the computer drew a card marked]* **Left**, what are your earnings (excluding the 3 tickets you receive for participating)?

_____ tickets

Your counterpart's earnings?

_____ tickets

2) Assume you stated that the minimum p for you to choose **In** over **Out** is 90% and p* is 50%. Further assume *your counterpart chose [the computer drew a card marked]* **Left**, what are your earnings (excluding the 3 tickets you receive for participating)?

_____ tickets

Your counterpart's earnings?

_____ tickets

3) The more unassigned tickets there are ...

- The higher my chances to win a prize.
- The higher my counterpart's chances to win a prize.
- The lower both my chances and my counterpart's chances to win a prize.

Decision situation *1 [2]*

For this situation:

- Your counterpart belongs to *your community [a different community]* and knows you are in *the same community [a different community]*.
- All Decisions B *[used to construct the deck of cards]* were made by members of *your community [other community/communities]*. When taking their decision, they knew that their counterpart was also *a member of the same community [from a different community]*.

KEY QUESTION: How large would the percentage p of *participants taking decision B who chose [cards marked]* Left minimally have to be for you to pick In over Out? (the number must lie between 0 and 100)

YOUR ANSWER: I choose In if p is at least _____

(this means that I choose Out if p is less than this cutoff)

□ ***Second movers in the trust game***

Quiz

1) If your counterpart chooses **In** and you choose **Right**, what are your earnings (on top of the 3 tickets you receive for participating)?

_____ tickets

Your counterpart's earnings?

_____ tickets

2) If your counterpart chooses **In** and you choose **Left**, what are your earnings (on top of the 3 tickets you receive for participating)?

_____ tickets

Your counterpart's earnings?

_____ tickets

3) If your counterpart chooses **Out** and you choose **Right**, what are your earnings (on top of the 3 tickets you receive for participating)?

_____ tickets

Your counterpart's earnings?

_____ tickets

4) If your counterpart chooses **Out** and you choose **Left**, what are your earnings (on top of the 3 tickets you receive for participating)?

_____ tickets

Your counterpart's earnings?

_____ tickets

5) The more unassigned tickets there are ...

- The higher my chances to win a prize.
- The higher my counterpart's chances to win a prize.
- The lower both my chances and my counterpart's chances to win a prize.

Decision situation *1 [2]*

For this situation, your counterpart is a member of *your community [a different community]* and knows when taking his/her decision that you are from *the same community [a different community]*.

KEY QUESTION: Which option, Left or Right, do you choose in case your counterpart chooses In?
YOUR ANSWER: I choose _____

- Left
- Right

□ ***All first movers and second movers in the trust game***
Survey

1. What is your gender?
    - Male
    - Female
    - Prefer not to answer

2. Please tell me, in general, how willing or unwilling you are to take risks. Please use a scale from 0 to 10, where 0 means you are "completely unwilling to take risks" and a 10 means you are "very willing to take risks".

completely unwilling to take risks										very willing to take risks
0	1	2	3	4	5	6	7	8	9	10

3. We now ask for your willingness to act in a certain way in four different areas. Please again indicate your answer on a scale from 0 to 10, where 0 means you are "completely unwilling to do so" and a 10 means you are "very willing to do so".

4. How well do the following statements describe you as a person? Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all" and a 10 means "describes me perfectly".

5. Assume you can give one extra lottery ticket to someone else. Who would you give it to?
    - A randomly chosen person from your community
    - A randomly chosen person who is taking *[the course in which students were recruited]*

	completely unwilling to do so										very willing to do so
	0	1	2	3	4	5	6	7	8	9	10
How willing are you to punish someone who treats **you** unfairly, even if there may be costs for you?	○	○	○	○	○	○	○	○	○	○	○
How willing are you to punish someone who treats **others** unfairly, even if there may be costs for you?	○	○	○	○	○	○	○	○	○	○	○

	does not describe me at all										describes me perfectly
	0	1	2	3	4	5	6	7	8	9	10
When someone does me a favor I am willing to return it.	○	○	○	○	○	○	○	○	○	○	○
I assume that people have only the best intentions.	○	○	○	○	○	○	○	○	○	○	○

2.C. Balancing tests

Table 2.4: Balancing tests: are first movers who answered the common comprehension questions correctly different?

	OLS (1)	OLS + time spent (2)	Session fixed effects (3)	Session fixed effects + time spent (4)	N
First movers in the trust game					
Active second mover	0.115	0.140	0.158	0.202	164
	(0.081)	(0.087)	(0.123)	(0.142)	
Risk aversion (0–10)	0.024	−0.027	−0.519	−0.537	164
	(0.298)	(0.279)	(0.412)	(0.398)	
Male	0.012	−0.028	−0.070	−0.153	163
	(0.099)	(0.094)	(0.132)	(0.120)	
First movers in the aligned interests game					
Active second mover	−0.025	−0.049	−0.031	−0.045	157
	(0.078)	(0.077)	(0.122)	(0.116)	
Risk aversion (0–10)	−0.402	−0.415	−0.427	−0.449	157
	(0.379)	(0.396)	(0.557)	(0.569)	
Male	0.019	−0.007	0.044	0.039	156
	(0.101)	(0.102)	(0.128)	(0.124)	
First movers with an active opponent					
Interests not aligned (TG versus AIG)	0.185 **	0.225 ***	0.179	0.237 *	161
	(0.070)	(0.076)	(0.114)	(0.123)	
Risk aversion (0–10)	−0.340	−0.354	−0.484	−0.528	161
	(0.316)	(0.295)	(0.421)	(0.392)	
Male	0.019	−0.015	0.034	−0.011	160
	(0.085)	(0.093)	(0.125)	(0.133)	
First movers with a passive opponent					
Interests not aligned (mTG versus mAIG)	0.046	0.031	0.079	0.055	160
	(0.078)	(0.076)	(0.125)	(0.124)	
Risk aversion (0–10)	0.001	−0.043	−0.371	−0.493	160
	(0.319)	(0.323)	(0.428)	(0.453)	
Male	0.009	−0.024	−0.029	−0.071	159
	(0.075)	(0.074)	(0.096)	(0.097)	

Notes: Each coefficient in (1)–(4) is from a separate regression. Each of the variables listed is regressed on a dummy for having answered the five comprehension questions common to all treatments correctly, with ((2), (4)) or without ((1), (3)) a control variable for the time spent reading the instructions in minutes.
We run regressions in four samples of first movers: in the trust game (TG or mTG), in the aligned interests game (AIG or mAIG), with an active opponent (TG or AIG), and with a passive opponent (mTG or mAIG). We report the coefficient of the dummy variable for having answered correctly. The figures in parentheses are standard errors robust to clustering at the session level. The sample sizes do not add up to those in Table 2.1 because due to an error 9 first movers in TG were not shown one of the comprehension questions. One participant did not answer the question about their sex. The p-values do not account for multiple comparisons.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

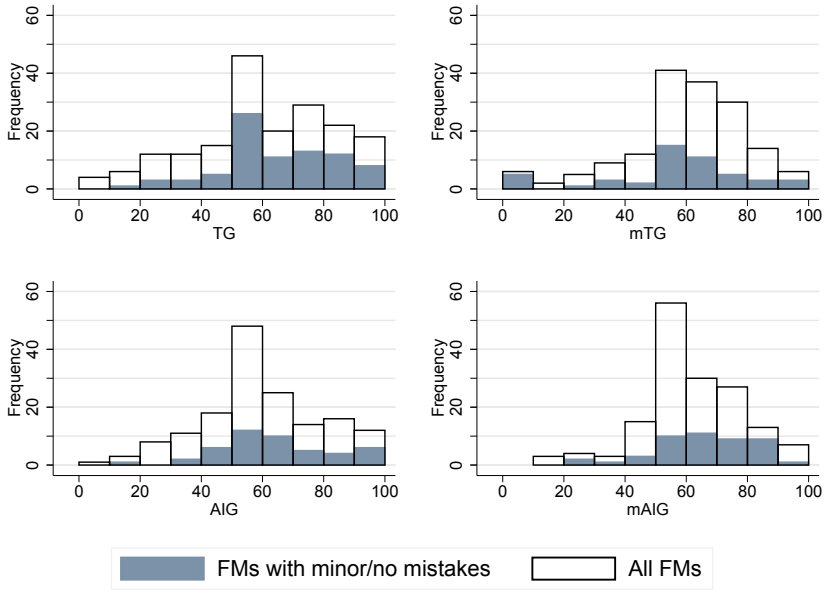


Figure 2.2: Distribution of Minimum Acceptable Probabilities across treatments
Notes: The graph shows the 660 MAPs of all 330 first movers (no-fill bars) versus the 222 MAPs of the 111 first movers who answered the understanding questions with minor or no mistakes (blue bars).

Chapter 3

Group identity and betrayal: decomposing trust

Abstract

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, 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.

I. 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; Chapter 2 in this dissertation; 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; Chapter 2 in

[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.

this dissertation). 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 three main 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 outcome-based social preferences (Cox, 2004; Ashraf et al., 2006; Sapienza et al., 2013). 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 Section II.C.

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).[2] 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 in- and outgroup members.

Second, in light of the null result at T2, where I did not find betrayal aver-

[2]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 3.A.

sion, 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.[3] 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; 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.[4] 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.[5]

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,

[3]The tweak to the original design in Bohnet and Zeckhauser (2004) and the issues it addresses are explained in detail in Section III.

[4]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, and is attributed to preferences (Becker, 2010).

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

[5]This identification strategy is related to identifying taste-based discrimination as a residual after controlling for beliefs about in- and outgroup members, which is widely used (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 III.

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 II presents the related literature. Section III describes the experimental design. Section IV explains the conceptual framework and presents the hypotheses. Section V presents the data and the results and Section VI concludes.

II. 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.[6] In this section, I first describe how betrayal aversion has been identified in Bohnet and Zeckhauser (2004) and 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.

II.A. 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 3.1. 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

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

trusting the second mover. The second mover's option *Left* is interpreted as her returning the first mover's trust.

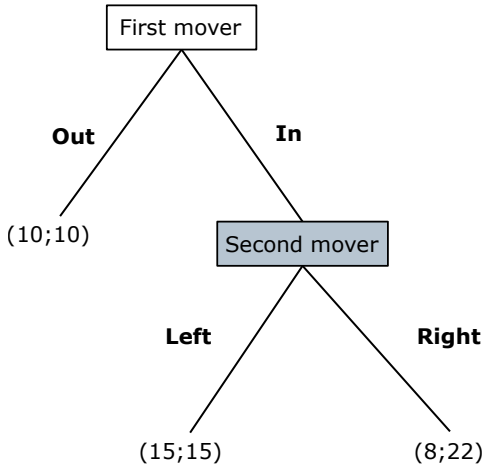


Figure 3.1: The binary trust game in Bohnet and Zeckhauser (2004)

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, 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.

II.B. 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). Chapter 2 in this dissertation replicates 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; Chapter 2 in this dissertation).[7] 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.

II.C. 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.[8] 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

[7] The data collected at T1 for this study is also part of a larger data set used in Chapter 2.

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

reflects discrimination in betrayal aversion or in one of the confounds pointed out by Li et al. (2020).

III. 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 3.2 presents the two treatments. Payoffs in Figure 3.2 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.[9]

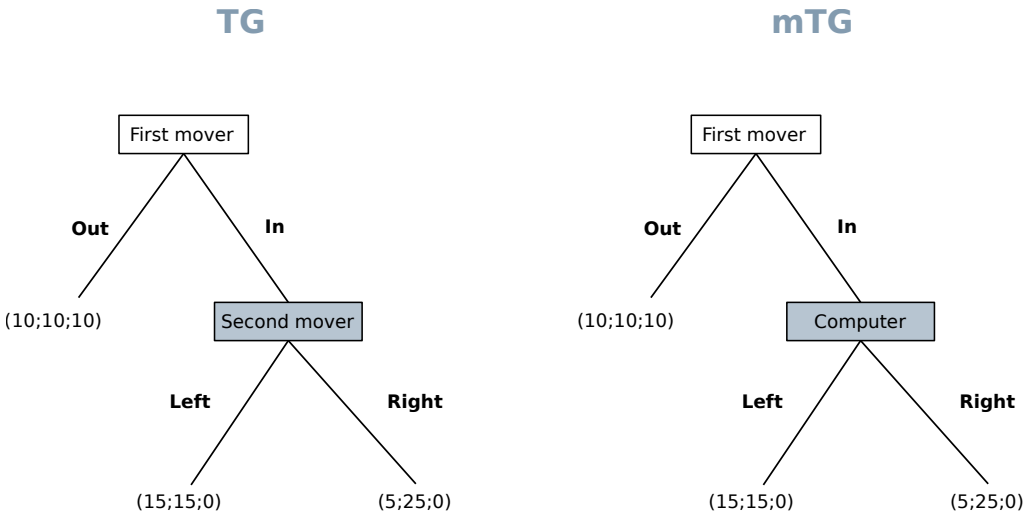


Figure 3.2: The treatments

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

[9]mTG is equivalent to BZ's Risky Dictator game. I refer to this treatment as "mTG" for consistency with the companion paper, which is Chapter 2 in this dissertation.

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 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.[10]

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 af-

[10]For details about the social groups and tests of the assumption that students perceived in- versus outgroup members differently see Appendix 3.A.

filiation". 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 discrimination: since these identities are not loaded with pre-existing stereotypes, differences in behavior towards in- an outgroup members are attributable exclusively to group membership. 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 means that in order to better understand discrimination as it is experienced with natural identities, it is useful to study discrimination with naturally occurring identities. The cleanest type of identification with natural identities is when the identities are randomly assigned (Götte et al., 2006). 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.

Table 3.1: 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. 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 3.1 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.[11] The experiment ended here for passive second movers. First movers and active second movers received instructions according to their treatment/role combination.

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.[12] This leads to only one third of first movers at T1 being included in the T1 estimation sample.[13]

At T2, after being able to use only one third of the data at T1, I decided to allow participants 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

[11] For details about the assignment to treatment and role and about the matching procedure, see Appendix 3.B.

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

[13] 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).

movers at T2 are included in the T2 estimation sample.[14]

Table 3.2 describes the samples at T1 and T2.

Table 3.2: 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.

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

[14] In Appendix 3.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.

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 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.[15] 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).[16]

The experiments were run in Qualtrics. The instructions for T1 are in Appendix 2.B of Chapter 2, as the trust game and modified trust game data at T1 are a subset of the data used in Chapter 2. The instructions at T2 were largely the same as at T1, with two major differences: (i) at T1, the experiment continued even if participants had not answered all comprehension questions correctly after two tries, while at T2, participants had to answer correctly to be allowed to continue, and (ii) participants were told that approximately 700 (600) students would take part at T1 (T2), as by T2 we had updated the attendance list in the study track by removing drop-outs.

IV. 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.

The hypotheses fall into two categories: those about behavior at T1 and T2,

[15]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.

[16]These numbers are higher than the totals in Table 3.2 as there were additional treatments, not discussed in this paper.

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.

IV.A. 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 that the willingness to accept the risky payoff from trusting ingroup members and that from trusting outgroup members do 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.

IV.B. 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 III 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.[17]

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 3.1 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). I expect beliefs about the opponent's trustworthiness at T2 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. This is similar to the result of Bacine and Eckel (2018).

IV.C. 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 IV.A and IV.B 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 (Hypothesis 7a) or less willing to trust outgroup members at T2 than at T1 (Hypothesis 7b),

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

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 (Hypothesis 8a) or less willing to enter the modified trust game with outgroup members at T2 than at T1 (Hypothesis 8b), or both.

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 (Hypothesis 9a), and betrayal aversion towards outgroup members does not decrease (Hypothesis 9b).

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.

V. Data and results

V.A. Summary statistics and nonparametric tests

Table 3.3 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 = 0.07), and also towards ingroup opponents (p-value = 0.07). 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 = 0.02, for all MAPs; p-value = 0.04, 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 = 0.83 at T1; p-value = 0.29 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 = 0.04 for $MAP_{mTG2,I}$ versus $MAP_{mTG1,I}$; p-value = 0.18 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 = 0.46 for $MAP_{TG2,I}$ versus $MAP_{TG1,I}$; p-value = 0.75 for $MAP_{TG2,O}$ versus $MAP_{TG1,O}$).[18]

V.B. 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 3.4) and T2 (in Table 3.5). In these regressions, I cluster errors at the individual level.

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

Table 3.3: Minimum acceptable probabilities

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

Table 3.4: 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 (**H2**)	4.58	4.58	3.18	3.18
	(3.69)	(3.72)	(4.08)	(4.13)
TG1 × Outgroup match (**H3**)	−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***	64.97***	47.44***	54.48***
	(5.19)	(9.02)	(7.91)	(12.36)
Linear combination				
TG1,I–TG1,O (**H1**)	0.89	0.89	−0.81	−0.81
	(3.35)	(3.38)	(3.29)	(3.33)
Session fixed effects			✓	✓
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 3.4, 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 following coefficients are important for hypotheses 1–3: the linear combination denoted in the table as "TG1,I–TG1, O" for H1, the coefficient of "Outgroup match" for H2, and the coefficient of the interaction term "TG1 $\times$ Outgroup match" for H3. 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. Since the coefficient of TG1 $\times$ Outgroup match is not significant in any specification, there is evidence in favor of H3: positive betrayal aversion which does not depend on the identity of the opponent. Playing with an in- as opposed to an outgroup opponent does not make a difference at T1 in either of the two treatments (in (4), p-value = 0.81 for the coefficient of TG1,I–TG1, O; p-value = 0.45 for the coefficient of Outgroup match). 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.[19,20]

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 no difference in the willingness to accept the risky payoff with externalities for an in- versus an outgroup opponent. There is also no difference in willingness to trust in- versus outgroup members.

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

[19] I nonetheless report results in columns (3) and (4), as results from the companion paper (Chapter 2 in this dissertation)—which includes additional treatments and thus has a bigger sample—confirm the sign and the significance level of betrayal aversion at T1.

[20] In the full sample (which includes first movers who did not answer the comprehension questions correctly) the coefficients of TG1, Outgroup match and their interaction have the same signs as in Table 3.4 and are insignificant. Results are available in Appendix Table 3.12.

Table 3.5: 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 (**H5**)	1.33	1.14	1.33	1.14
	(2.13)	(2.18)	(2.19)	(2.24)
TG2 × Outgroup match (**H6**)	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***	83.68***	67.17***	86.18***
	(4.20)	(7.78)	(9.28)	(10.91)
Linear combination				
TG2,I–TG2,O (**H4**)	−5.28*	−5.28*	−5.28*	−5.28*
	(2.86)	(2.87)	(2.93)	(2.95)
Session fixed effects			✓	✓
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 3.5, I repeat the same exercise for the data collected at T2. The following coefficients are important for hypotheses 4–6: the linear combination denoted in the table as "TG2,I–TG2, O" for H4, the coefficient of "Outgroup match" for H5, and the coefficient of the interaction term "TG2 $\times$ Outgroup match" for H6. Here, in all four specifications, the coefficient quantifying betrayal aversion towards the ingroup is negative and insignificant. In column (4), the coefficient of TG2,I–TG2, O is negative and marginally significant, offering moderate support for H4 (p-value = 0.08). The coefficients corresponding to H5 and H6 are insignificant in all specifications.

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 running the regressions in Table 3.5 separately for first decisions (in Table 3.6) and for second decisions (in Appendix Table 3.13). For first decisions, I find even stronger support for H4: the coefficient of TG2,I–TG2, O is negative and has p-value = 0.04 in specification (4). There is no support for H5 and H6 in the sample of first decisions. All relevant coefficients are insignificant in the sample of second decisions (see Appendix Table 3.13).

Table 3.6: Linear regressions on Minimum Acceptable Probabilities at T2: first decision

	(1)	(2)	(3)	(4)
Baseline: mTG2, ingroup match				
TG2	−6.49	−5.95	−8.63	−8.27
	(6.48)	(6.30)	(6.20)	(6.20)
Outgroup match (**H5**)	2.94	5.91	4.89	6.97
	(6.48)	(6.39)	(6.26)	(6.25)
TG2 × Outgroup match (**H6**)	8.54	6.36	10.42	8.23
	(9.59)	(9.24)	(9.75)	(9.53)
Risk loving (0–10)		−3.09***		−2.69**
		(1.09)		(1.20)
Male		−0.85		−1.49
		(5.09)		(5.53)
Constant	60.81***	78.80***	61.86***	76.87***
	(4.27)	(7.59)	(9.98)	(11.01)
Linear combination				
TG2,I–TG2,O (**H4**)	−11.47	−12.27*	−15.30**	−15.20**
	(7.06)	(6.71)	(7.45)	(7.14)
Session fixed effects			✓	✓
Adjusted R^2	0.00	0.06	0.03	0.08
Observations	93	92	93	92
Individuals	93	92	93	92
Sessions	10	10	10	10

Notes: Robust standard errors 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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.[21]

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 IV, I look into heterogeneous effects at T2 by the strength of the attachment to the ingroup as proxied by the hypothetical allocation task. In Appendix Table 3.14, I run regressions from Table 3.5 on the subsample of first decisions of participants who select an ingroup recipient in the hypothetical allocation task (56 first movers). While these participants are not betrayal averse to neither in- nor outgroup members, they are willing to trust ingroup members for lower rates of trustworthiness (p-value = 0.02 for the coefficient of TG2,I–TG2,O in (4)). They are also more willing to take a risky bet with externalities for another participant if this participant is an ingroup member rather than an outgroup member (p-value = 0.02

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

for the coefficient of Outgroup match in (4)).

In Appendix Table 3.15, I do the same on the subsample of first decisions of participants who select a random recipient from the entire subject pool in the allocation task (37 first movers). None of hypotheses 4–6 are supported in this sample, but these results should be interpreted with caution as the sample is small.[22]

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 the ticket to a random recipient from the entire subject pool do not state different MAPs in in- versus outgroup matches. Neither of the two types of first movers displays betrayal aversion on average, neither towards in- nor towards outgroup opponents.

V.C. 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

[22] I decided to split the sample of first decisions in two groups depending on the recipient selected for the hypothetical allocation task as the alternative would have been to have a triple interaction term (between the treatment dummy, the opponent identity dummy and the dummy for selecting an ingroup recipient) in a bigger sample. The results of the analysis in this bigger sample with the triple interaction are qualitatively similar to those in Appendix Tables 3.14 and 3.15.

[23] In the pooled data on both decisions at T2, only the coefficient of TG2,I–TG2,O is significant (and negative) among those selecting an ingroup recipient in the hypothetical allocation task (p-value = 0.05).

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 3.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: 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 3.7 below—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.

I find strong support for H9a: 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)). In model (4), H9b is contradicted: there is a significant decrease in betrayal aversion towards outgroup members between T1 and T2 as well, which runs counter to the expected increase (p-value = 0.03). Also in (4), there is marginally significant evidence supporting H7a: first movers ask for lower MAPs in the modified trust game with ingroup members at T2 than at T1 (p-value = 0.09).[24] Even though not significant, the signs of the coefficients corresponding to H7a and H7b are negative in all specifications, whereas those corresponding to hypotheses H8a and H8b are almost always positive. This suggests that decreases in betrayal aversion between T1 and T2 happen because

[24]While it remains negative, this coefficient turns insignificant if I use other sessions as baseline. The coefficients corresponding to H9a and H9b are not sensitive to changing the baseline session.

of a lowering of the threshold to trust both in- and outgroup members and because of an increase (in seven out of eight specifications, all eight insignificant) in the threshold to take a risky gamble with an in- and an outgroup member. The decrease in betrayal aversion for ingroup members during this period does not differ from the decrease in betrayal aversion for outgroup members (none of the coefficients corresponding to H10 is significant).

Result 5 Betrayal aversion towards ingroup members decreases significantly between T1 and T2, and so does that towards outgroup members. The two decreases do not differ significantly from each other.

In conclusion, I observe decreases in betrayal aversion which I cannot attribute to increases or decreases in the blocks that make up betrayal aversion, as the coefficients of these blocks are insignificant.

Table 3.7: 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)
O	4.58	4.58	3.18	3.18
	(3.65)	(3.67)	(3.81)	(3.84)
TG × O	−5.48	−5.48	−2.37	−2.37
	(4.94)	(4.96)	(4.90)	(4.93)
T2 (**H8a**)	10.47*	8.78	4.60	1.14
	(5.95)	(5.78)	(11.95)	(11.01)
TG × T2 (**H9a**)	−16.94**	−15.73**	−20.03***	−22.52***
	(7.60)	(7.56)	(7.22)	(7.15)
O × T2	−3.26	−3.45	−1.86	−2.04
	(4.23)	(4.27)	(4.42)	(4.47)
TG × O × T2 (**H10**)	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***	71.00***	58.60***	82.88***
	(4.94)	(6.88)	(7.77)	(9.74)
Linear combinations				
T2 + TG × T2 (**H7a**)	−6.47	−6.95	−15.43	−21.37*
	(4.72)	(4.80)	(13.35)	(12.43)
T2 + TG × T2 + O × T2 + TG × O × T2 (**H7b**)	−0.29	−0.77	−10.96	−16.90
	(5.13)	(4.99)	(13.44)	(12.43)
T2 + O × T2 (**H8b**)	7.21	5.34	2.74	−0.90
	(5.89)	(5.83)	(12.27)	(11.37)
TG × T2 + TG × O × T2 (**H9b**)	−7.51	−6.11	−13.70*	−16.00**
	(7.81)	(7.73)	(7.44)	(7.25)
Session fixed effects			✓	✓
Adjusted R^2	0.00	0.05	0.07	0.14
Observations	316	314	290	288
Individuals	158	157	145	144
Sessions	43	43	30	30

Notes: Standard errors clustered at the individual level in parentheses. "TG1" refers to TG at T1, "mTG1" to mTG at T1, etc. "O" stands for outgroup match. 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).

VI. 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; Chapter 2 in this dissertation) 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 first movers make in the second experiment, there is ingroup bias in conditional trust.

When looking more closely at this first decision, 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. For those who select an ingroup recipient, social distance affects both trusting decisions and 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. The reasoning is that these components are what the two treatments have in common (see Table 3.1).

In neither of the two experiments is there discrimination in betrayal aversion. Participants become less betrayal averse towards both in- and outgroup members as time passes, with betrayal aversion being insignificantly different from zero in the second experiment. This could mean that the social identity manipulation did not work as expected, as no bias of favoring the in- over the outgroup emerged over time, neither in trust nor in the component capturing beliefs and outcome-based social preferences. It could also mean that familiarity and frequent interaction outweighed social group concerns for the decisions in the experiment on average. Social distance matters for a small majority of players, in that it influences a component aggregating their risk preferences, beliefs about trustworthiness and outcome-based social preferences. Yet another possibility for the lack of ingroup bias on average is that subjects become better over time in understanding statistics, and that once one controls for subjective beliefs in such a sample, there is no betrayal aversion (so no difference in betrayal aversion due to mechanical reasons). Future research should check whether betrayal aversion survives controlling for subjective beliefs in other contexts, which do not necessarily involve social identity.

The lack of betrayal aversion in the second experiment could also be due to other reasons, such as 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), it is possible they were more likely to make more analytical, "System 2" (Kahneman, 2003) 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, and that in such a sample, there is no betrayal aversion after controlling for beliefs, as explained above. For instance, 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 no longer 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.

Appendix

3.A. The social groups: creating the ingroup/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 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.[25]

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

[25]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.

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 3.8.

After the first two months of studying students were asked to evaluate their community's functioning.[26] A higher number indicates more agreement with a statement. Statement 3 (in bold) 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 3.8: 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 exper-

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

iments 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, while 60% do so at T2. A two-tailed Mann-Whitney test shows this difference is not significant (p-value = 0.43).

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 0.54). However, at T2 this difference is significant at 5% (p-value = 0.05).[27]

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

[27]The results of chi-square tests on intergroup discrimination are similar for the sample of active second movers (p-value = 0.27 at T1—I only consider second movers with correct answers to the understanding questions; p-value = 0.04 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 < 0.01 at T2), while they do not at T1 (from T1, only including those with correct answers: p-value = 0.26). This suggests that ingroup favoritism in altruism may have developed by T2.

3.B. Assignment to treatment and matching procedure

Note: *This section is very similar to the one described in Appendix 2.A of Chapter 2 in this dissertation. 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.[28] 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*;[29]

– 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.

[28]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.

[29]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.

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 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.[30]

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.[31]

[30] 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.

[31] 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.

3.C. Balancing tests and robustness check

A balancing test in Appendix Table 3.9 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 3.9: 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 3.10, 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 questions correctly, but it includes 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 3.2), 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.[32] 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

[32]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 3.10: 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 ***	−0.387 ***
	(0.066)	(0.056)
Time spent reading instructions (min)	−0.805 **	−1.664 ***
	(0.342)	(0.432)
Total duration (min)	−1.047 *	−0.634
	(0.625)	(0.663)
Was in mTG	0.037	0.136 **
	(0.037)	(0.059)
Male	−0.083	−0.070
	(0.063)	(0.082)
Risk loving (0–10)	−0.293	−0.271
	(0.274)	(0.325)
Others can be trusted	0.073	0.093
	(0.295)	(0.330)
Positively reciprocal	−0.128	−0.201
	(0.158)	(0.166)
Negatively reciprocal if treated unfairly	−0.015	−0.042
	(0.295)	(0.397)
Negatively reciprocal if others treated unfairly	−0.134	−0.552
	(0.320)	(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.

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).

Table 3.11: Minimum acceptable probabilities at T2

	TG	
	Correct answers	Incorrect answers
	58.19	56.90
	(27.28)	(25.14)
p-value	0.667	
Observations	62	30
Individuals	31	15
	mTG	
	Correct answers	Incorrect answers
	62.97	60.51
	(23.74)	(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.

Table 3.12: Linear regressions on Minimum Acceptable Probabilities at T1: full sample

	(1)	(2)	(3)	(4)
Baseline: mTG1, ingroup match				
TG1	1.06	0.43	2.00	1.52
	(3.12)	(3.11)	(3.28)	(3.29)
Outgroup match (**H2**)	1.35	1.35	1.38	1.38
	(1.89)	(1.90)	(2.06)	(2.07)
TG1 × Outgroup match (**H3**)	−1.96	−1.96	−1.56	−1.56
	(3.26)	(3.27)	(3.50)	(3.51)
Risk loving (0–10)		−2.15**		−1.53**
		(0.70)		(0.72)
Male		2.05		3.61
		(2.88)		(2.94)
Ingroup first	−1.22	−0.92	−1.05	−0.66
	(2.81)	(2.77)	(2.95)	(2.90)
Constant	56.41***	69.42***	61.73***	66.54***
	(2.45)	(4.86)	(5.23)	(6.62)
Linear combination				
TG1,I-TG1,O (**H1**)	0.62	0.62	0.18	0.18
	(2.65)	(2.66)	(2.82)	(2.83)
Session fixed effects			✓	✓
Adjusted R^2	−0.01	0.02	0.06	0.07
Observations	346	346	340	340
Individuals	173	173	170	170
Sessions	45	45	42	42

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 all first movers at T1. The sample in models (3) and (4) consists of those first movers at T1 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.D. Checking for order effects

Table 3.13: Linear regressions on Minimum Acceptable Probabilities at T2: second decision

	(1)	(2)	(3)	(4)
Baseline: mTG2, ingroup match				
TG2	−5.00	−6.23	−5.02	−7.13
	(7.75)	(7.99)	(7.42)	(7.46)
Outgroup match (**H5**)	−0.35	−3.45	−2.53	−5.26
	(6.87)	(7.13)	(6.93)	(6.79)
TG2 × Outgroup match (**H6**)	−0.98	0.94	−3.45	−0.55
	(10.91)	(10.77)	(10.72)	(10.52)
Risk loving (0–10)		−3.30**		−3.26**
		(1.30)		(1.43)
Male		1.12		−1.71
		(5.76)		(5.49)
Constant	60.88***	82.14***	66.65***	87.26***
	(5.11)	(10.44)	(9.70)	(12.68)
Linear combination				
TG2,I–TG2,O (**H4**)	1.33	2.51	5.98	5.81
	(8.47)	(8.06)	(8.24)	(8.07)
Session fixed effects			✓	✓
Adjusted R^2	−0.02	0.02	0.06	0.11
Observations	93	92	93	92
Individuals	93	92	93	92
Sessions	10	10	10	10

Notes: Robust standard errors 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.14: Linear regressions on Minimum Acceptable Probabilities at T2, first decision: hypothetical ticket to ingroup member

	(1)	(2)	(3)	(4)
Baseline: mTG2, ingroup match				
TG2	−0.70	−2.00	−2.02	−3.83
	(8.18)	(7.95)	(8.63)	(9.25)
Outgroup match (**H5**)	15.10*	16.82**	16.99*	18.01**
	(7.77)	(8.27)	(8.47)	(8.76)
TG2 × Outgroup match (**H6**)	−0.15	−1.65	1.21	0.16
	(10.89)	(10.64)	(11.17)	(11.35)
Risk loving (0–10)		−2.73**		−1.57
		(1.21)		(1.57)
Male		0.43		−2.48
		(5.70)		(6.45)
Constant	57.98***	74.06***	72.98***	81.87***
	(5.68)	(10.11)	(8.12)	(11.12)
Linear combination				
TG2,I–TG2,O (**H4**)	−14.95*	−15.17**	−18.20**	−18.18**
	(7.63)	(6.93)	(7.99)	(7.68)
Session fixed effects			✓	✓
Adjusted R^2	0.07	0.11	0.15	0.15
Observations	56	55	56	55
Individuals	56	55	56	55
Sessions	10	10	10	10

Notes: Robust standard errors in parentheses. "TG2" refers to TG at T2, "mTG2" to mTG at T2, etc. 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.15: Linear regressions on Minimum Acceptable Probabilities at T2, first decision: hypothetical ticket to any student in course

	(1)	(2)	(3)	(4)
Baseline: mTG2, ingroup match				
TG2	−18.16*	−13.99	−14.91	−13.52
	(10.15)	(11.69)	(11.72)	(13.20)
Outgroup match (**H5**)	−14.56	−9.60	−12.22	−11.35
	(9.16)	(8.52)	(10.39)	(10.83)
TG2 × Outgroup match (**H6**)	23.62	19.71	20.58	19.68
	(15.77)	(16.02)	(16.00)	(17.03)
Risk loving (0–10)		−2.65		−0.95
		(2.67)		(2.63)
Male		−2.48		−0.10
		(10.52)		(12.33)
Constant	67.29***	81.27***	47.72***	53.24**
	(5.15)	(13.74)	(15.35)	(24.01)
Linear combination				
TG2,I–TG2,O (**H4**)	−9.06	−10.11	−8.36	−8.33
	(12.83)	(13.53)	(13.01)	(13.86)
Session fixed effects			✓	✓
Adjusted R^2	−0.02	−0.03	0.02	−0.06
Observations	37	37	37	37
Individuals	37	37	37	37
Sessions	9	9	9	9

Notes: Robust standard errors in parentheses. "TG2" refers to TG at T2, "mTG2" to mTG at T2, etc. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.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.[33] 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.[34] In both models, I cluster standard errors at the counterfactual

[33] 642 students registered for the first exam session in the study track from which I recruited participants.

[34] In the model with session fixed effects, the size of the counterfactual estimation samples at

individual level.

Table 3.16 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 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 3.7 tell the same story about the change in behavior between T1 and T2.

Table 3.16: 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.

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.

Chapter 4

Testing the elicitation of the Minimum Acceptable Probability

Abstract

Betrayal aversion has been shown to be an important determinant of trust (Bohnet and Zeckhauser, 2004). We study whether the way betrayal aversion is identified (as a difference in Minimum Acceptable Probabilities, MAPs) is affected by beliefs about one's prospects.

In a within-subject design, we find that MAPs are lower the worse the prospects one faces. This is similar to the distributional dependence of valuations elicited using the Becker–DeGroot–Marschak mechanism. Our results suggest that distributional dependence should be accounted for when eliciting MAPs to isolate betrayal aversion.

This chapter is co-authored with Martin Strobel.

I. Introduction

Individuals have often been found to prefer exposure to a randomly generated risk over exposure to an equiprobable risk generated by an opponent in a strategic situation. In the context of trust games, this strategic risk premium has been dubbed *betrayal aversion* (Bohnet and Zeckhauser, 2004). Many papers find that betrayal aversion is an important determinant of trust (Bohnet and Zeckhauser, 2004; Aimone et al., 2015; Fairley et al., 2016; Quercia, 2016; Bacine and Eckel, 2018; Butler and Miller, 2018; Chapter 2 in this dissertation).

Betrayal aversion is identified as the difference in first mover behavior in two games: a binary trust game—a version of the trust game (Berg et al., 1995)—and an equivalent game where the decision at the second node is made by a randomization device. In both games, first movers have to decide whether to keep their endowment or to send it to the second mover. If the first mover sends money, there is an efficiency gain. The second mover (either a real decision maker or a randomization device) decides whether to share the gain fairly or to keep most of it.

Typically first movers do not decide directly, but indicate their minimum acceptable probability (MAP). This is the lowest probability for which first movers prefer sending money over keeping their endowment. After all relevant second movers' decisions are collected, the actual probability of a fair split is calculated over the entire pool of second mover decisions. Then, the first mover sends the money if the actual probability is larger or equal to their minimum acceptable probability. If he does, the payoffs are decided by a randomly assigned second mover's decision.

The mechanism resembles the Becker–DeGroot–Marschak mechanism (Becker et al., 1964, in short, BDM). The MAP is elicited without first movers knowing how many second movers (devices) chose the favorable outcome at the second node. It is in first movers' best interest to state the true MAP at which they prefer sending money over not sending it.

For expected utility maximizers, the MAP should be independent of their belief about the actual probability of fair sharing. This need not be the case for non-expected utility maximizers. A recent paper shows theoretically that the elicitation procedure of MAPs used in most papers on betrayal aversion leaves

the door open to potential confounds for betrayal aversion such as "ambiguity attitudes, complexity, different beliefs, and dynamic optimization" if players are not rational expected utility maximizers (Li et al., 2020). Moreover, a couple of empirical papers which use more stringent identification procedures for betrayal aversion by controlling for first mover beliefs in the two games do not find betrayal aversion (Fetchenhauer and Dunning, 2012, the second experiment in Chapter 3), or find it to play a role for trusting only when beliefs are far more optimistic than is generally the case (Engelmann et al., 2021).

In this note, we use an online experiment to measure how much of what has been called betrayal aversion is due to distributional dependence, regardless of the source of risk being random or strategic. We remove the strategic component and show participants complete distributions over probabilities of the good (and bad) outcome of a lottery, and ask them to state their MAP for preferring the lottery over a safe payoff. When deciding about the MAP, participants do not know which lottery will be relevant, but they know the distribution from which the lottery will be drawn. Some refer to such situations as involving ambiguity, others—as involving complex risks (the compound risk of which lottery will be selected and what the outcome of the lottery will be). In this paper, we refer to the situation as involving complex risk.[1]

Following Li et al. (2020), we expect a premium between the distribution mimicking the control condition in betrayal aversion studies and the distribution mimicking the binary trust game condition. We find the opposite to be true: the higher the expected probability of the favorable outcome, the higher the minimum acceptable probability required by participants to prefer the lottery.

While this is at odds with our expectations, it ties in with findings from the empirical literature on distributional dependence of willingness to pay (WTP) as elicited through the BDM mechanism. Similarly to betrayal aversion, theoretical literature has pointed out that the BDM mechanism is not incentive compatible if players are not rational expected utility maximizers (Karni and Safra, 1987; Horowitz, 2006). This is because individuals face uncertainty regarding the price of the good at stake and additional uncertainty about whether

[1]Ambiguity aversion and attitudes to complex risks are positively correlated (Armantier and Treich, 2016).

they will buy the good or not. If their utility function is influenced by these uncertainties, changing the price distribution of the good might influence their valuation of the good (here, the MAP). Several empirical papers find this to be the case for the BDM: generally, the higher the expected price of the good, the higher the WTP (for a short review of this literature, see Tymula et al., 2016). The results of Tymula et al. (2016) are partly consistent with theories of reference-dependent preferences (Kőszegi and Rabin, 2006, 2007; Wenner, 2015).

Our results suggest that (i) the way MAPs are elicited is sensitive to subjective beliefs, so these should be taken into account in order not to confound valuation, and (ii) the way subjective beliefs influence valuation is not in line with results of the toy model in Li et al. (2020).

The paper is structured as follows. Section II describes the experimental design and procedures. Section III sets forth the hypothesis. Section IV presents the results. Section V explains how our results inform the existing literature and suggests directions for future research.

II. Design and procedures

We use a within-subject design, with each subject being exposed to all treatments sequentially. In each treatment, participants see a graphical representation of a distribution over lotteries with two possible outcomes (high and low), but varying probabilities for each outcome. A lottery will be drawn at random from the distribution. This means in some treatments it is more likely to get a lottery with a high chance of a high payoff than in others. We use three distributions over lotteries. The distributions are ordered in terms of the expected payoff over the entire distribution, as their name suggests: the Good, the Bad, and the Uniform (the Good $>$ the Uniform $>$ the Bad).

To make the task easy to understand, we present lotteries via 32 wheels of fortune with 15 sectors each. Dark blue sectors symbolize the high payoff (£4), light blue sectors—the low payoff (£1). The sure payoff (the payoff participants receive if no wheel is spun) is £2. In each treatment, participants see the wheels sorted in ascending order by the probability of the favorable outcome, with the 32 wheels equally distributed over 4 rows. Figure 4.1 below shows the distribution of lotteries for the Good treatment.

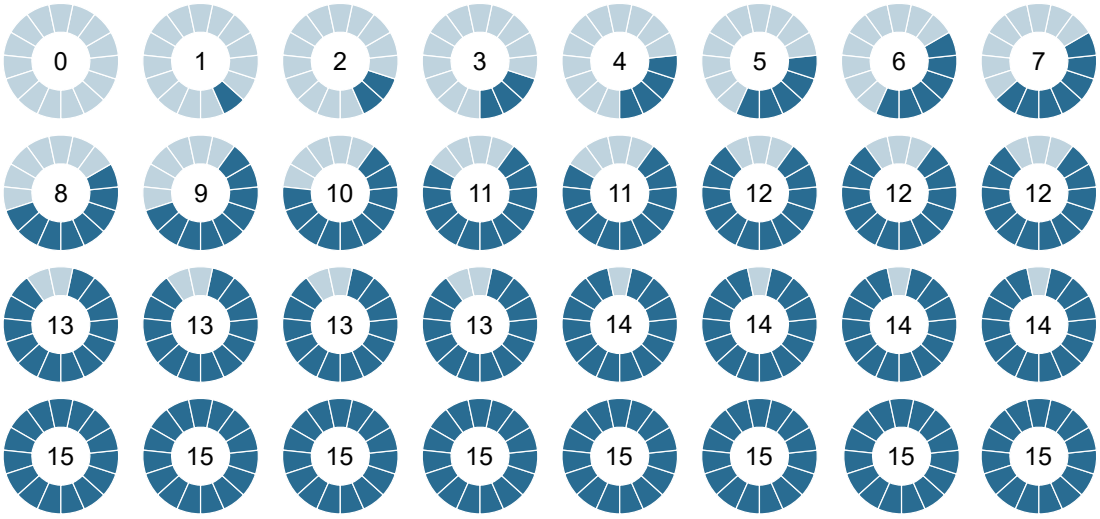


Figure 4.1: The Good distribution

Two of the three distributions are meant to emulate treatments in papers on betrayal aversion. The Uniform distribution has equal chances of occurrence for each of the possible wheels. We assume that this is what participants expect to face in treatments with decisions made by randomization devices, unless specified otherwise.[2] The Bad distribution has an overall chance of a high payoff similar to the share of trustworthy respondents in Western samples in papers on betrayal aversion (0.2895) (e.g. Bohnet and Zeckhauser, 2004; Bohnet et al., 2008). The distribution in Good mirrors the one in Bad: its overall expected chance of a high payoff is one minus that in Bad (0.7105), it has the same variance and minus the skewness of the Bad distribution. We included this distribution to check if departures from the Uniform distribution in either direction yield effects of similar size (albeit reverse sign) on reported MAPs. Table 4.1 presents the distributions.

Participants are told that one of the wheels will be drawn at random, with all wheels having an equal chance to be drawn. They are asked to state a *minimum acceptable frequency*: the lowest number of dark blue sectors in the randomly drawn wheel such that they prefer to spin the wheel instead of receiving the sure payoff.[3] Specifically, they have to answer: "Which wheels would you like

[2]We assume participants in the *Risky Dictator Game* in Bohnet and Zeckhauser (2004) had such a distribution in mind.

[3]We decided to use frequencies instead of probabilities because there is evidence that participants have an easier time expressing choice this way (Quercia, 2016).

Table 4.1: The treatments: the distribution of chances of a high payoff

	# of wheels		
# of high payoff sectors	The Good	The Bad	The Uniform
0	1	8	2
1	1	4	2
2	1	4	2
3	1	3	2
4	1	2	2
5	1	1	2
6	1	1	2
7	1	1	2
8	1	1	2
9	1	1	2
10	1	1	2
11	2	1	2
12	3	1	2
13	4	1	2
14	4	1	2
15	8	1	2
Total # of wheels	32	32	32

to spin for your bonus?" by inserting an integer between 0 and 15 in the blank space: "I prefer to spin wheels which have at least ____ dark blue sectors."[4]

The experiment was conducted online using Qualtrics. Participants were UK residents registered on a platform for conducting academic studies (Prolific). Since the elicitation of MAPs is rather complex (Quercia, 2016; Chapter 3 in this dissertation), we opted for participants who had at least a bachelor's degree. The study was pre-registered at the AEA RCT Registry (`https://doi.org/10.1257/rct.7776-1.1`).

The study had three stages: a set of eliminatory comprehension questions, the three decisions, and a post-experimental questionnaire.[5] Those who com-

[4]We chose the setup with wheels of fortune as we wanted to make the task easy to understand. Despite our approach being discrete, we will interpret the frequencies (x out of 15) as minimum acceptable probabilities. Some papers on betrayal aversion also use a discrete approach by asking respondents how many second movers from the pool of possible matches should reciprocate for them to prefer sending money (e.g. one of the experiments in Quercia, 2016).

[5]In the post-experimental questionnaire, respondents answered an unincentivized question to determine their ambiguity aversion, a version of a cognitive reflection test (Frederick, 2005; Thomson and Oppenheimer, 2016) adapted by the authors, a question about the subject they

pleted the experiment (went only through the comprehension questions) spent a median time of 12.4 (5.9) minutes and earned on average 3.96 (1) UK pounds.[6]

We present the instructions in Appendix 4.A.

III. Hypothesis

Let p be the probability of the high payoff and $1-p$ the probability of the low payoff of the lottery. The distribution of p (and consequently, of $1-p$) varies between treatments. Based on Li et al. (2020) we assume that what has been called betrayal aversion could be due to such differences in the underlying distribution of p.

Specifically, we adapt the toy example in Appendix A in Li et al. (2020) to predict the optimal MAP in each treatment. We additionally assume that participants treat complex bets similarly to how they treat ambiguous bets (for supporting evidence, see Armantier and Treich, 2016). This leads us to expect the following ordering of MAPs:[7]

Hypothesis 1 *The MAP in Good (more mass on high values of p) is lower than the MAP in Uniform (a uniform distribution over p), which is lower than the MAP in Bad (more mass on low values of p).*

$$MAP_G < MAP_U < MAP_B \tag{4.1}$$

We also consider the alternative hypothesis ($MAP_B < MAP_U < MAP_G$).

studied for their most recent degree, a general risk taking question (Dohmen et al., 2011), a question about their aspiration level for earnings from participating in a survey, a couple of questions to check their anchoring susceptibility, from which an anchoring score can be computed (Cheek and Norem, 2017), a set of questions about their optimism/pessimism, the revised Life Orientation Test (Scheier et al., 1994) and a brief sensation seeking scale, BSSS-4 (Stephenson et al., 2003).

[6]Participants were paid £1 for going through the comprehension questions (regardless of the correctness of their answers). Those who answered the comprehension questions correctly earned an additional £1, £2 or £4 for one of their decisions.

The high average earnings of those who completed the experiment are due to a coding error which we detected after running the experiment. Instead of decisions in all three treatments being equally likely to be selected, only those in Good and Uniform were selected, each with equal probability. This led all participants who had completed all stages of the experiment to have a higher chance of a higher payoff. This error did not affect decisions, but only which decision was selected for payment. Participants were informed about the error after the experiment.

[7]For details, see Appendix 4.B.E.

This could be true if participants anchor their MAPs on visual or numerical cues of the distributions, such as the mean.

IV. Results

IV.A. The estimation sample

Table 4.2 describes the sample. Treatment was assigned in order to balance the number of participants exposed to each of the six possible orderings of treatments. 275 of the 450 participants answered the eliminatory comprehension questions correctly and completed the experiment. Since assignment to treatment happened before participants had gone through the comprehension questions, this leads to slightly different sizes of the subsamples for the six orderings.

Table 4.2: Characteristics of the estimation sample

	Age	Share male	Sample size
Good–Uniform–Bad	30.956	0.333	45
	(8.808)	(0.477)	
Uniform–Bad–Good	33.538	0.346	52
	(9.074)	(0.480)	
Bad–Good–Uniform	37.114	0.523	44
	(11.071)	(0.505)	
Good–Bad–Uniform	33.132	0.491	53
	(9.174)	(0.505)	
Bad–Uniform–Good	32.429	0.333	42
	(9.423)	(0.477)	
Uniform–Good–Bad	33.333	0.205	39
	(10.103)	(0.409)	
Total	33.411	0.378	275
	(9.685)	(0.486)	

Notes: The table shows averages per sequence. Standard deviations in parentheses.

IV.B. Behavior in the experiment

First, we present summary statistics for all decisions, by treatment and by decision order. Next, we run two-sided nonparametric tests and ordinary least

squares regressions to test the hypothesis.

Table 4.3 presents the average MAP by treatment over all decisions and by decision order. This table already suggests that the hypothesis is not supported by the data, as the average MAP is highest in Good, followed by Uniform, followed by Bad (except for the second decision).

Table 4.3: Descriptive statistics: MAPs by treatment (x out of 15)

	All decisions	First decision	Second decision	Third decision
The Good	9.531	9.571	9.458	9.553
	(2.503)	(2.270)	(2.500)	(2.750)
The Uniform	8.844	8.890	8.368	9.227
	(2.382)	(2.392)	(2.119)	(2.539)
The Bad	8.615	8.093	9.124	8.512
	(2.522)	(2.597)	(2.491)	(2.387)
N	825	275	275	275

Notes: The table shows averages per treatment. Each participant made three decisions in randomized order. Standard deviations in parentheses. Possible answers were integers between 0 and 15.

A nonparametric Page's L test confirms this: there is strong evidence that the ordering is the opposite to the one hypothesized ($MAP_B < MAP_U < MAP_G$, p-value < 0.001).[8]

Figure 4.2 shows that this ordering of MAPs holds for all six sequences. One sequence stands out: Good–Bad–Uniform. For each treatment, MAPs in this sequence are higher than in any other sequence. Even in the first decision, the MAP in **Good**–Bad–Uniform differs significantly from its counterpart in **Good**–Uniform–Bad (p-value = 0.003, Mann-Whitney test). Since the sequence of events and the information participants faced up to that point in the two sequences had been identical, this difference cannot be a treatment effect, nor an order effect.

In Table 4.4 we present results of ordinary least square regressions of MAPs. Model (1) contains as regressors only dummy variables indicating the treatment. Model (2) adds age and gender as explanatory variables. Model (3) additionally includes risk attitudes. Model (4) also includes dummy variables

[8]Page's L test has the null hypothesis that all possible orderings are equally likely. The alternative hypothesis is that a specified order is the increasing order of alternatives. The Stata command is $pagetrend$.

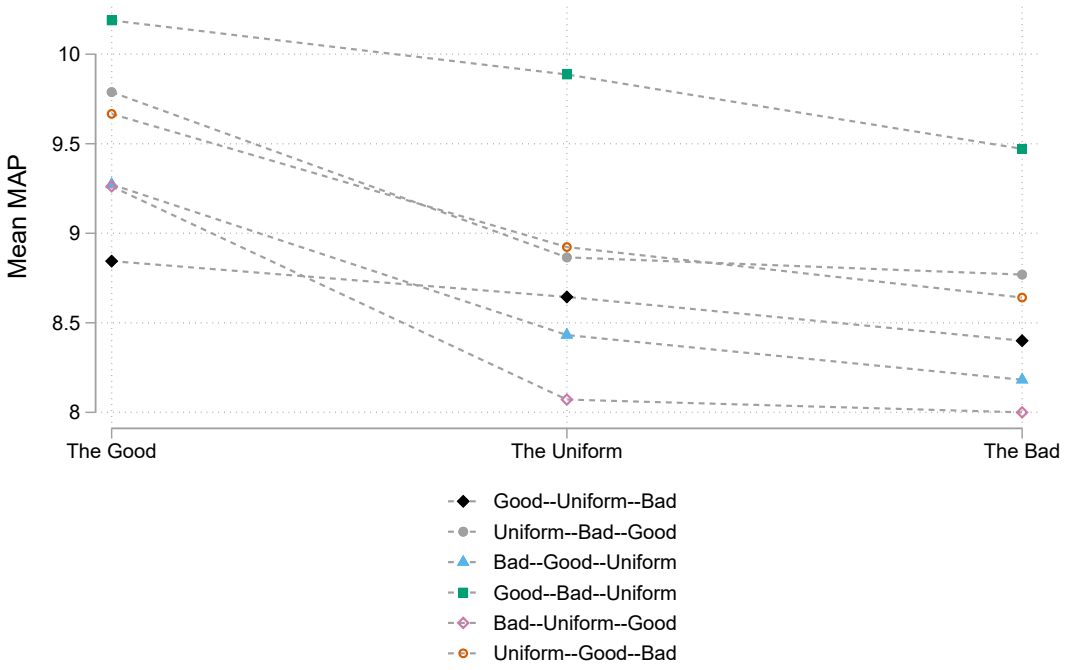


Figure 4.2: Mean MAPs by treatment and decision sequence

for the order in which participants were exposed to treatments. In all models, standard errors are clustered at the individual level.

In all four specifications, participants ask for 0.687 more dark blue sectors (yielding a high payoff) on average in Good compared to Uniform to be willing to spin the selected wheel (p-value < 0.001 in all specifications). They also ask for 0.229 fewer dark blue sectors in Bad compared to Uniform (p-value $= 0.001$ in (4)). More risk loving individuals have lower MAPs (p-value $= 0.04$ in (4)).[9,10,11]

[9]We used ordinary least squares regressions for ease of interpretation of the coefficients. Since the dependent variable is categorical and ordered, we also used ordered logit models. The results are qualitatively similar. Compared to Uniform, MAPs between 1 and 8 are less likely in Good (more likely in Bad) and MAPs between 9 and 15 are more likely in Good (less likely in Bad).

[10]In a robustness check, we reran the regressions separately for each ordering. The signs of the effects are the same for each ordering as they are in the pooled sample, though some effects do not reach significance in these smaller samples.

[11]The coefficient of the sequence which stands out in Figure 4.2, Good–Bad–Uniform, is not significant with the baseline treatment and baseline sequence used in model (4). We also ran a contrast analysis after this regression, to check how the coefficients of each sequence differ from the grand mean. Even after using a Bonferroni correction, the coefficient of this sequence

Table 4.4: Linear regressions on Minimum Acceptable Frequencies

Dependent variable:	Minimum acceptable frequency			
	(1)	(2)	(3)	(4)
The Good	0.687 ***	0.687 ***	0.687 ***	0.687 ***
	(0.099)	(0.099)	(0.099)	(0.099)
The Bad	−0.229 ***	−0.229 ***	−0.229 ***	−0.229 ***
	(0.070)	(0.070)	(0.070)	(0.070)
Age		0.005	0.006	0.006
		(0.013)	(0.013)	(0.014)
Male		−0.047	0.030	−0.029
		(0.286)	(0.284)	(0.283)
Risk aversion (0–10)			−0.172 **	−0.152 **
			(0.074)	(0.074)
Sequence				
Good–Uniform–Bad				−0.490
				(0.408)
Bad–Good–Uniform				−0.497
				(0.475)
Good–Bad–Uniform				0.683
				(0.461)
Bad–Uniform–Good				−0.640
				(0.491)
Uniform–Good–Bad				−0.020
				(0.489)
Constant	8.844 ***	8.696 ***	9.520 ***	9.556 ***
	(0.144)	(0.460)	(0.593)	(0.696)
N	825	825	825	825

Notes: Standard errors clustered at the individual level in parentheses. The baseline treatment is the Uniform distribution. The baseline sequence in (4) is Uniform–Good–Bad. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Coefficients of The Good and The Bad differ between models, but only in the fourth or higher decimal. This is also true for the standard errors.

Result 1 *Participants set the lowest requirement to be willing to take a randomly drawn lottery in Bad, followed by Uniform, followed by Good.*

Subjects' MAPs are stickier if they start with Good than with the other two: the intra-individual standard deviation over all three MAPs is lower if the first decision is in Good than if it is in one of the other two treatments (Mann-Whitney test, p-value = 0.02). Table 4.5 shows the results of running specifications (1)–(3) in Table 4.4 on first decisions only. Since the skewed effect of stickiness is not present, deviations in MAP in Good and in Bad do not differ in absolute size (Wald test for equality of coefficients in (3), p-value = 0.87). The smaller coefficient in Bad over all three decisions is thus due to more pronounced stickiness when facing prospects that worsen than when facing prospects that improve over time.

Result 2 *Within individual, MAPs are stickier for participants who face the Good first than for those who face one of the other two distributions first.*

Table 4.5: Linear regressions on Minimum Acceptable Probabilities: first decision (x out of 15)

Dependent variable:	Minimum acceptable frequency		
	(1)	(2)	(3)
The Good	0.681 *	0.731 **	0.682 *
	(0.352)	(0.355)	(0.353)
The Bad	−0.797 **	−0.794 **	−0.783 **
	(0.363)	(0.367)	(0.364)
Age		0.018	0.019
		(0.015)	(0.015)
Male		−0.194	−0.113
		(0.303)	(0.303)
Risk attitudes (0–10)			−0.174 **
			(0.076)
Constant	8.890 ***	8.333 ***	9.189 ***
	(0.253)	(0.573)	(0.680)
N	275	275	275

Notes: The baseline treatment is the Uniform distribution. 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.

is the only one which is significantly higher than the grand mean. As mentioned before, this is a particularity of the data which cannot be attributed to treatment effects, nor to order effects.

We speculated that such an ordering of MAPs is possible if individuals anchor on visual or numerical cues offered by the distributions. If this were true, then the effects should be reduced if we add an interaction term between the individual anchoring score (Cheek and Norem, 2017) as measured in the post-experimental questionnaire and the treatments. This is however not the case: if we include the interaction term in models in Table 4.4, the coefficients of The Good and The Bad keep their magnitude and significance levels. Treatment effects do not differ for those who are more or less susceptible to anchoring. Someone who is one standard deviation less susceptible to anchoring than the mean (in either direction) asks for a MAP which is higher by approximately 0.58 (significant at 10% level, results available on request).

A suggestion we received after the data collection was that instead of thinking in terms of MAPs, subjects might be attracted to the visual center of the distributions.[12] Should this be the case, the ordering of MAPs would coincide with the one we observe for a mechanical reason. In order to test this, we rerun the specifications in Table 4.4, but we use as dependent variable the number of wheels which, if randomly selected, are relevant for the participant's payoff. In other words, this is the number of wheels which—given the participant's MAP—if selected, would be spun. We consider this assumption to be supported if either (i) treatment does not influence the number of wheels potentially spun and this number is close to 16 in all treatments (half of the 32 wheels available) or (ii) treatment influences significantly the number of wheels potentially spun, but the coefficients of the treatment variables are small in a "real-world" sense.

Table 4.6 shows that in Uniform, approximately 14 wheels are potentially spun for payoff on average. While this is close to the expected 16 wheels, the number of wheels varies significantly for the other two treatments. In Good, subjects are willing to spin approximately 7.4 more wheels—the equivalent of an additional row of wheels. In Bad, subjects are willing to spin approximately 6.8 fewer wheels.[13] We conclude that while there is a potential "pull towards the visual center" effect, it cannot explain the results.

Another suggestion was that our results could be explained by the range-frequency theory (Parducci, 1965; Parducci and Perrett, 1971).[14] This theory

[12]We thank Mats Köster for this suggestion.
[13]The results are similar for the sample of first decisions.
[14]We thank Andrea Isoni for this suggestion.

Table 4.6: Linear regressions on wheels potentially spun for payoff

Dependent variable:	Wheels potentially spun for payoff			
	(1)	(2)	(3)	(4)
The Good	7.378 ***	7.378 ***	7.378 ***	7.378 ***
	(0.192)	(0.192)	(0.192)	(0.193)
The Bad	−6.785 ***	−6.785 ***	−6.785 ***	−6.785 ***
	(0.159)	(0.159)	(0.159)	(0.160)
Age		−0.009	−0.010	−0.011
		(0.021)	(0.021)	(0.022)
Male		0.166	0.057	0.147
		(0.462)	(0.462)	(0.459)
Risk attitudes (0–10)			0.242 *	0.209 *
			(0.123)	(0.123)
Sequence				
Good–Uniform–Bad				0.736
				(0.653)
Uniform–Bad–Good				−0.104
				(0.768)
Bad–Good–Uniform				0.820
				(0.778)
Good–Bad–Uniform				−1.185
				(0.796)
Bad–Uniform–Good				0.875
				(0.821)
Constant	14.313 ***	14.539 ***	13.380 ***	13.415 ***
	(0.288)	(0.745)	(0.986)	(1.173)
N	825	825	825	825

Notes: Standard errors clustered at the individual level in parentheses. The baseline treatment is the Uniform distribution. The baseline sequence in (4) is Uniform–Good–Bad. Risk attitudes are measured on a 0–10 scale, where 0 is very risk averse and 10 is very risk loving. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
The standard errors are not constant across specifications, but they differ in the fourth or higher decimal.

states that when presented with stimuli (physical, such as sounds or weights, but also monetary rewards), participants bin the stimuli into categories depending on the available range of stimuli and on their frequency. Participants do this as a compromise between (i) dividing the available range into equal shares and (ii) ensuring that each bin has an equal share of stimuli. If participants consider that only certain categories are acceptable risks, this can lead to the MAP ordering observed in our data. We provide a numerical example of such a rationale in Appendix 4.C.

V. Discussion

In this note, we test a necessary assumption used in the way betrayal aversion has been elicited in the past. This assumption is that the underlying distribution of probabilities does not matter for the choice of MAP. If the underlying distribution does matter, then betrayal aversion is misidentified.

We remove the social/strategic aspects of the original game and exogenously manipulate underlying distributions in three treatments. Two of these treatments aim to emulate plausible distributions imagined by subjects in studies on betrayal aversion. We find a difference in behavior between treatments, but opposite to our expectation: the more favorable the distribution of lotteries, the better the lotteries have to be to be preferred to the safe option. We thus find a distributional dependence of risk attitudes as elicited using MAPs, but of the opposite sign than the predictions of the toy example in Li et al. (2020). This result implies that betrayal aversion should be identified after controlling for subjective beliefs.

The result is consistent with several theories. A first type of such theories are theories of reference-dependent preferences which predict that individuals will be more risk loving when endowed with riskier options. Since our experiment was not meant to disentangle between competing theories, several of them could explain our results—for instance Kőszegi and Rabin (2006, 2007) or Wenner (2015). These theories state that expectations (which we manipulated exogenously by changing the underlying distribution of lotteries) act as reference points. Modifying expectations modifies the gain-loss component of the utility function, such that higher expectations may make the same outcome less desirable. Alternatively, changing expectations could directly affect con-

sumption utility: if one derives self-image utility from one's consumption, a change in expectations could change which goods are more desirable and thus, which ones offer a boost in self-image for the owner (Strahilevitz and Loewenstein, 1998; Marzilli Ericson and Fuster, 2011). With better options overall, the bar to determine which of them increase one's status is placed higher. A second theory which could explain the result is the range-frequency model (Parducci, 1965; Parducci and Perrett, 1971). This theory explicitly considers that when evaluating the intensity of a stimulus, participants take into account both the range and the frequency of available stimuli. They divide the stimuli into categories according to each criterion and populate the categories with a roughly equal number of observations. By changing the frequency, as we do in the treatments, we change which components a participant assigns to a category. If only certain categories are deemed acceptable (i.e. risks worth taking), this can affect decisions in a way which aligns with the results.

We chose the distributions for the treatments such that the overall chance of a high payoff was close to the probability of trustworthiness in the original studies on betrayal aversion. Further decisions about the Bad (Good) distribution were based on the condition that optimal MAPs be different in the three treatments using the parameters in the toy example of Li et al. (2020) and additional assumptions detailed in Appendix 4.B.E. Many distributions fit this criterion and our choice at this point was arbitrary. Our results point to the need to account for beliefs in the control treatment used to identify betrayal aversion (the Risky Dictator Game in Bohnet and Zeckhauser, 2004). Future evidence on how people think about random versus strategic risk and ambiguity will hopefully reconcile results from betrayal aversion studies with those on the flexible valuation of risky goods.

Appendix

4.A. Instructions[15]

Statement of consent

In this study, you will be asked to make decisions. You will also be asked to answer comprehension questions, reasoning questions, and questions about yourself. Your data will remain anonymous in accordance with GDPR (the European Union's personal data protection law).

This study follows the guidelines of the BEELab at Maastricht University. This means that all information you receive during the study is truthful.

To continue, please select "I agree to participate".

- I agree to participate
- I don't agree to participate

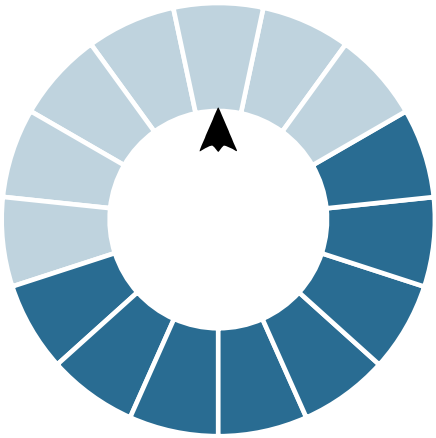
Before you start, please switch off your phone/e-mail/music so you can focus on this study. Thank you!
Please enter your Prolific ID: _____

Part 1
This part explains what the study is about and presents examples. We will test your understanding of the situation with some questions.

[15]While the instructions look slightly differently in Qualtrics, we strove to depict the visual elements (the wheels, the distributions) as accurately as possible in this section. The horizontal lines mark the separation between pages.

To continue to Part 2, you have to answer these questions correctly.

Consider a wheel of fortune like the one below. The wheel is equally likely to land on each sector. The pointer indicates the result: it's the sector which ends up at 12 o'clock when the wheel stops spinning. Give it a try!



Spin the wheel!

In Part 2, you will see more such wheels. All wheels have **15 sectors** in total, which are either light blue or dark blue. The number in the middle is the **number of dark blue sectors** in a wheel.
Below is an example with five wheels.

One of the wheels will be randomly selected. Each wheel is **equally likely** to be selected. If the selected wheel is spun, it is equally likely to land on each sector.
You will have the following options for your bonus:
Let us consider some examples. If the selected wheel has

DON'T SPIN	You don't spin the selected wheel. Your bonus is £2.00.
SPIN	You spin the selected wheel. Your bonus is £4.00 if the wheel lands on dark blue, and £1.00 if it lands on light blue.

- 15 dark blue sectors, if you **SPIN** it your bonus is £4.00 for sure. If you **DON'T SPIN** it, you are guaranteed to receive £2.00.
- 0 dark blue sectors, if you **DON'T SPIN** it you are guaranteed £2.00. If you **SPIN** it, your bonus is £1.00 for sure.

Without knowing which wheel has been selected, you will be asked which wheels you want to **SPIN** for your bonus, and which ones you **DON'T** want to **SPIN**.

You will be asked the following question:

Which wheels do you prefer to SPIN for your bonus?

I prefer to SPIN wheels which have at least [ ] *dark blue sectors.*

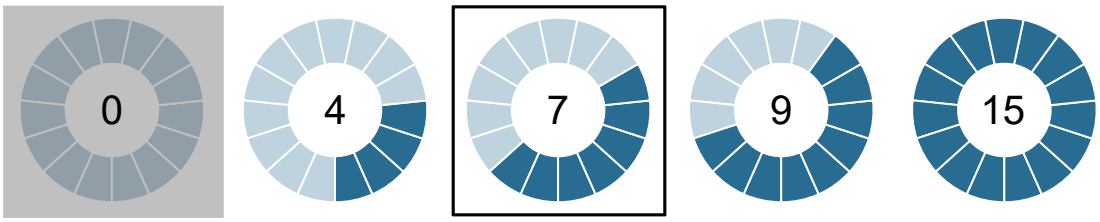
*If the randomly selected wheel has **fewer than ... dark blue sectors**, I **DON'T SPIN** it. I receive £2.00.*
*If the randomly selected wheel has **... or more dark blue sectors**, I **SPIN** it. I receive £4.00 if the wheel lands on dark blue, and £1.00 if it lands on light blue.*

You can practice by introducing integers between 0 and 15 in the box above. When you introduce a number, all wheels with fewer dark blue sectors than your answer will be grayed out, indicating that you prefer DON'T SPIN for those wheels. The wheels with the same number or more dark blue sectors than your answer will not be affected, indicating that you prefer SPIN if one of those wheels is selected.
At the end of the study you will be told which wheel has been selected. Its number of dark blue sectors will be compared to your answer, and your bonus will be determined by the relevant option (SPIN or DON'T SPIN).
Your **input on this screen** is simply for you to practice and it **doesn't affect your bonus**. You **don't have to memorize** this explanation: a non-interactive

version like the one linked below under “View explanation” will be available whenever relevant.
View explanation

The comprehension questions will start on the next screen.

Comprehension Question 1



Consider the wheels above. Let’s assume you stated that you want to SPIN wheels with at least 3 sectors for your bonus. For this reason, wheels with fewer than 3 sectors are grayed out. The wheel with 7 sectors has been randomly selected (the wheel with a black border).

Please select **all that apply**.

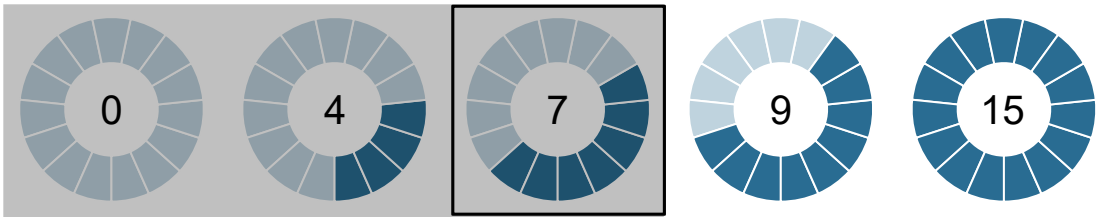
View explanation[16]

- ☐ I DON’T SPIN the selected wheel.
- ☐ I SPIN the selected wheel.
- ☐ My bonus is £1.00.
- ☐ My bonus is £2.00.
- ☐ My bonus is £4.00 if the selected wheel lands on dark blue, £1.00 if it lands on light blue.
- ☐ My bonus is £1.00 if the selected wheel lands on dark blue, £2.00 if it lands on light blue.

---

[16]Upon clicking, a pdf document opened in a separate window. The document contained the text on page 180 (“Consider a wheel...”) up to page 181 (“... determined by the relevant option (SPIN or DON’T SPIN).”).

Comprehension Question 2



Consider the wheels above. Let's assume you stated that you want to SPIN wheels with at least 8 sectors for your bonus. For this reason, wheels with fewer than 8 sectors are grayed out. The wheel with 7 sectors has been randomly selected (the wheel with a black border).

Please select **all that apply**.

View explanation[17]

- ☐ I DON'T SPIN the selected wheel.
- ☐ I SPIN the selected wheel.
- ☐ My bonus is £1.00.
- ☐ My bonus is £2.00.
- ☐ My bonus is £4.00 if the selected wheel lands on dark blue, £1.00 if it lands on light blue.
- ☐ My bonus is £1.00 if the selected wheel lands on dark blue, £2.00 if it lands on light blue.

---

[17]Upon clicking, a pdf document opened in a separate window. The document contained the text on page 180 ("Consider a wheel...") up to page 181 ("... determined by the relevant option (SPIN or DON'T SPIN).").

Comprehension Question 3[18]

Please select the correct statement from each of the following pairs.

- Each wheel is **equally likely** to be selected.
- I **can** influence the chance that a particular wheel is selected.
- I get to spin the selected wheel **regardless of** whether it is in the grayed out area or not.
- If I get to spin the selected wheel, it is **equally likely** to land on each sector.
- My bonus is £4.00 if the selected wheel is in the **grayed out** area and the selected wheel lands on **light blue**.

- Some wheels are **more likely** to be selected than others.
- I **can't** influence the chance that a particular wheel is selected.
- I get to spin the selected wheel **only if** it's not in the grayed out area.
- If I get to spin the selected wheel, it is **more likely** to land on sectors which are initially around 12 o'clock.
- My bonus is £4.00 if the selected wheel is in the **non-grayed out** area and the selected wheel lands on **dark blue**.

You have answered all questions in Part 1 correctly.
You will now be directed to Part 2.

Part 2[19]

In this part, you will be asked

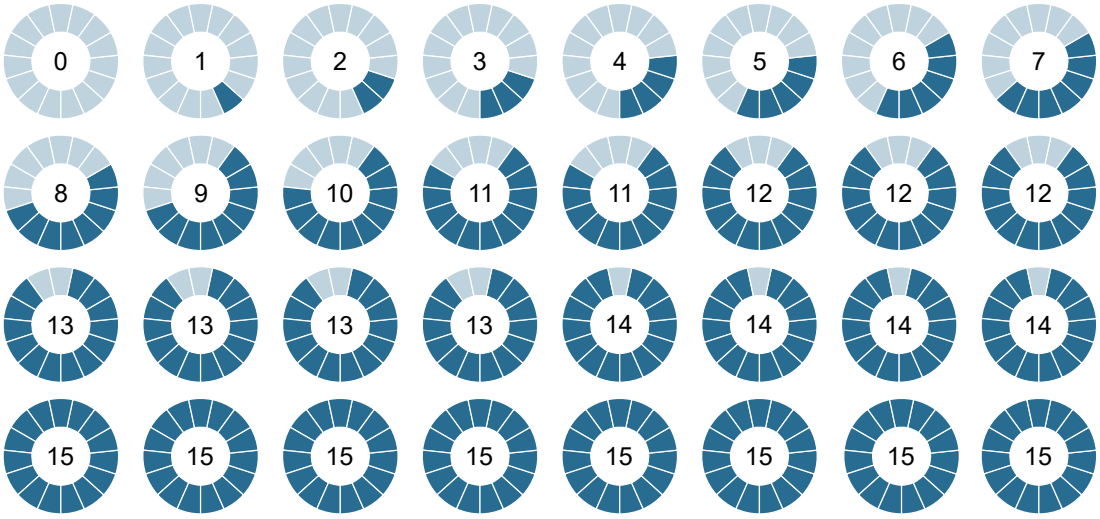
- How you want your bonus to be determined **in three different situations**. Choose your most preferred option from those available. There are no right or wrong answers to these questions.
- Reasoning questions and questions about yourself.

At the end of Part 2, **one of the three situations will be randomly selected**, and your bonus will be determined according to your answer in that situation. Each of the three situations is equally likely to be selected.

---

[18] If participants answered all comprehension questions correctly, they would go to the next part. If not, they were given the chance to review their answers. Those who revised correctly would also go to the next part. The experiment ended for those who did not revise correctly.

[19] The decisions on the next three screens were shown in randomized order.



Consider the wheels above. Which wheels do you prefer to SPIN for your bonus?

Please enter an integer between 0 and 15.

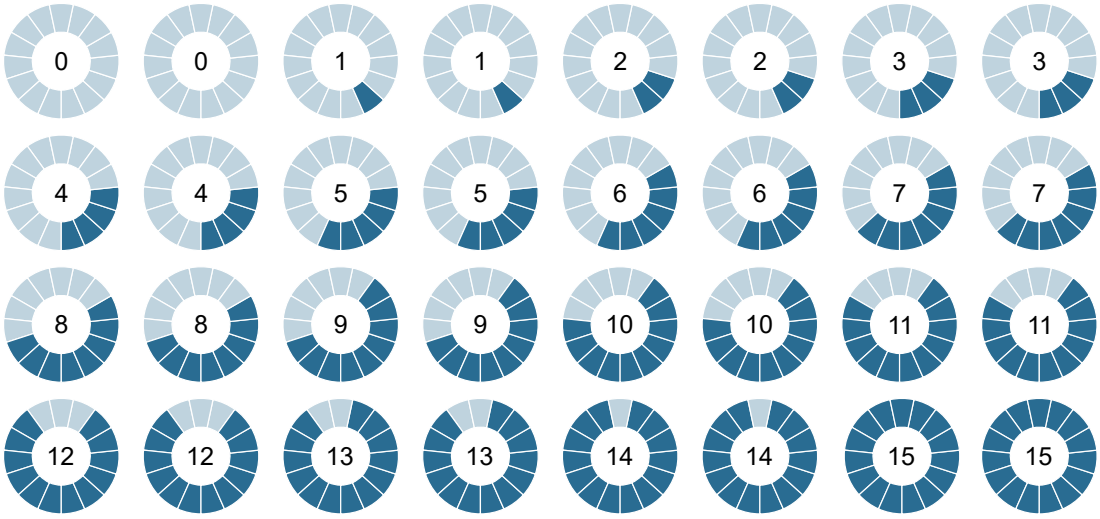
I prefer to SPIN wheels which have at least [ ][20] dark blue sectors.

If the randomly selected wheel has fewer than ... dark blue sectors, I DON'T SPIN it. My bonus is £2.00.
If the randomly selected wheel has ... or more dark blue sectors, I SPIN it. My bonus is

- £1.00 if the selected wheel lands on light blue, and
- £4.00 if it lands on dark blue.

---

[20]The box was dynamic: as participants typed a number, the wheels which were ineligible for being selected in case that number was the participant's decision were grayed out and the ellipses below were replaced with that number. This way, participants were informed about the implications of potential decisions.



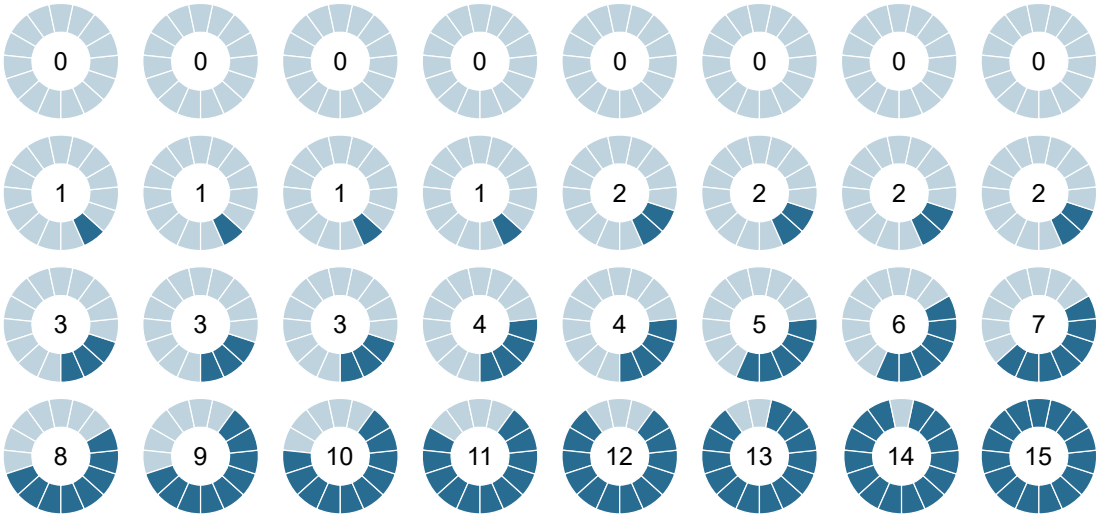
Consider the wheels above. Which wheels do you prefer to SPIN for your bonus?

Please enter an integer between 0 and 15.

I prefer to SPIN wheels which have at least ☐ dark blue sectors.

If the randomly selected wheel has fewer than ... dark blue sectors, I DON'T SPIN it. My bonus is £2.00.
If the randomly selected wheel has ... or more dark blue sectors, I SPIN it. My bonus is

- £1.00 if the selected wheel lands on light blue, and
- £4.00 if it lands on dark blue.



Consider the wheels above. Which wheels do you prefer to SPIN for your bonus?

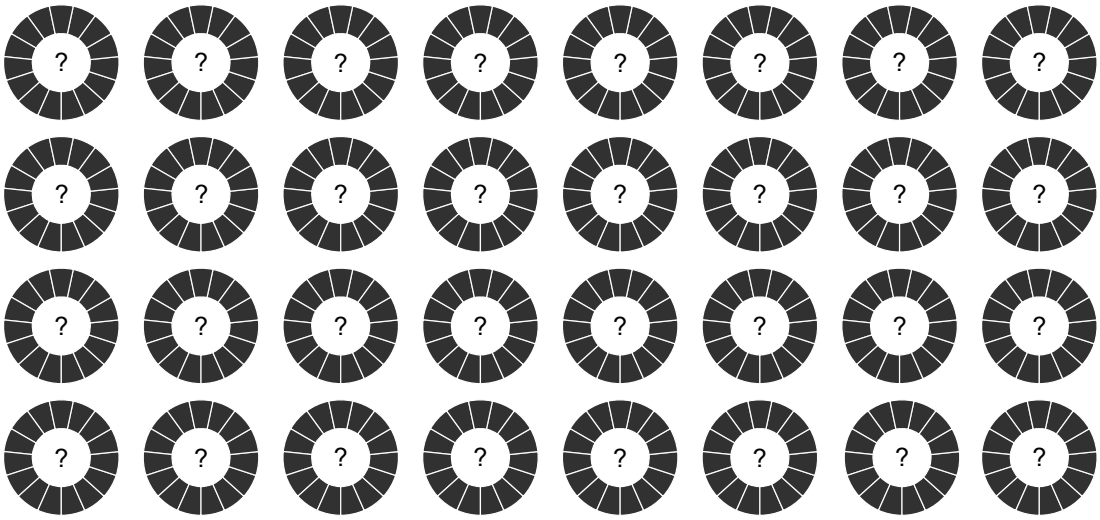
Please enter an integer between 0 and 15.

I prefer to SPIN wheels which have at least [ ] dark blue sectors.

If the randomly selected wheel has fewer than ... dark blue sectors, I DON'T SPIN it. My bonus is £2.00.
If the randomly selected wheel has ... or more dark blue sectors, I SPIN it. My bonus is

- £1.00 if the selected wheel lands on light blue, and
- £4.00 if it lands on dark blue.

The reasoning questions and the questions about yourself will start on the next screen.



The scenario described below is **hypothetical**: your answer **doesn't influence your bonus**. Imagine 32 wheels of fortune with 15 sectors each. Like before, their sectors are either light blue (worth £1.00) or dark blue (worth £4.00). However, the sectors' color is hidden, so you **don't know how many dark blue** or **light blue sectors** each wheel has. One of the 32 wheels will be selected at random: depending on your answer to the question below, you SPIN the wheel or you DON'T SPIN it.

In this case, which wheels would you prefer to SPIN for your bonus?

Please enter an integer between 0 and 15.

I prefer to SPIN wheels which have at least [ ] dark blue sectors.

If the randomly selected wheel has fewer than ... dark blue sectors, I DON'T SPIN it. My bonus is £2.00.

If the randomly selected wheel has ... or more dark blue sectors, I SPIN it. My bonus is

- £1.00 if the selected wheel lands on light blue, and
- £4.00 if it lands on dark blue.

Please answer the following questions.

Simon had 17 plants at home and all but 8 died. How many are left?

Claire's grandmother has three granddaughters. The first two are named April and May. What is the third granddaughter's name?

If you're running a race and you pass the person in second place, what place are you in? (type in the number of the place)

A scientist grows bacteria on a Petri dish. Every day, the area covered by bacteria doubles in size. If it takes 6 days for the entire dish to be covered, how long would it take for half of the dish to be covered?

What subject did you study for your most recent degree? *[drop-down menu]*

How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?

completely unwilling to take risks										very willing to take risks
0	1	2	3	4	5	6	7	8	9	10

Which of the following do you take into consideration when deciding whether to take part in a study? Please select all that apply.[21]

- Total pay
- Pay per hour
- Other things, such as

[21] One of the next two screens was shown at random to participants.

The next questions are about general facts that you may or may not know. Please give your best estimates. We also ask that you please not look up the answers; we are interested in people's estimates, whether or not they are accurate.

Do you think that the **average daily temperature in June** in Amsterdam, the Netherlands, between 1971 and 2020 was higher or lower than 14°C?

- Higher
- Lower

What do you think was the **average daily temperature in June** in Amsterdam in this period?

[ ] °C

Do you think that the number of **average daily hours of sunshine in June** in Amsterdam, the Netherlands, between 1971 and 2020 was higher or lower than 10?

- Higher
- Lower

What do you think was the number of **average daily hours of sunshine in June** in Amsterdam in this period?

[ ] hour(s) and [ ] minute(s)

The next questions are about general facts that you may or may not know. Please give your best estimates. We also ask that you please not look up the answers; we are interested in people's estimates, whether or not they are accurate.

Do you think that the **average daily temperature in June** in Amsterdam, the Netherlands, between 1971 and 2020 was higher or lower than 17°C?

- Higher
- Lower

What do you think was the **average daily temperature in June** in Amsterdam in this period?

[ ] °C

Do you think that the number of **average daily hours of sunshine in June** in Amsterdam, the Netherlands, between 1971 and 2020 was higher or lower than 4?

- Higher
- Lower

What do you think was the number of **average daily hours of sunshine in June** in Amsterdam in this period?

[ ] hour(s) and [ ] minute(s)

For the questions below, please be as honest and accurate as you can throughout. Try not to let your response to one statement influence your responses to other statements. There are no "correct" or "incorrect" answers. Answer according to your own feelings, rather than how you think "most people" would answer.

	I disagree a lot	I disagree a little	I neither agree nor disagree	I agree a little	I agree a lot
In uncertain times, I usually expect the best.	○	○	○	○	○
It's easy for me to relax.	○	○	○	○	○
If something can go wrong for me, it will.	○	○	○	○	○
I'm always optimistic about my future.	○	○	○	○	○
I enjoy my friends a lot.	○	○	○	○	○

	I disagree a lot	I disagree a little	I neither agree nor disagree	I agree a little	I agree a lot
It's important for me to keep busy.	○	○	○	○	○
I hardly ever expect things to go my way.	○	○	○	○	○
I don't get upset too easily.	○	○	○	○	○
I rarely count on good things happening to me.	○	○	○	○	○
Overall, I expect more good things to happen to me than bad.	○	○	○	○	○

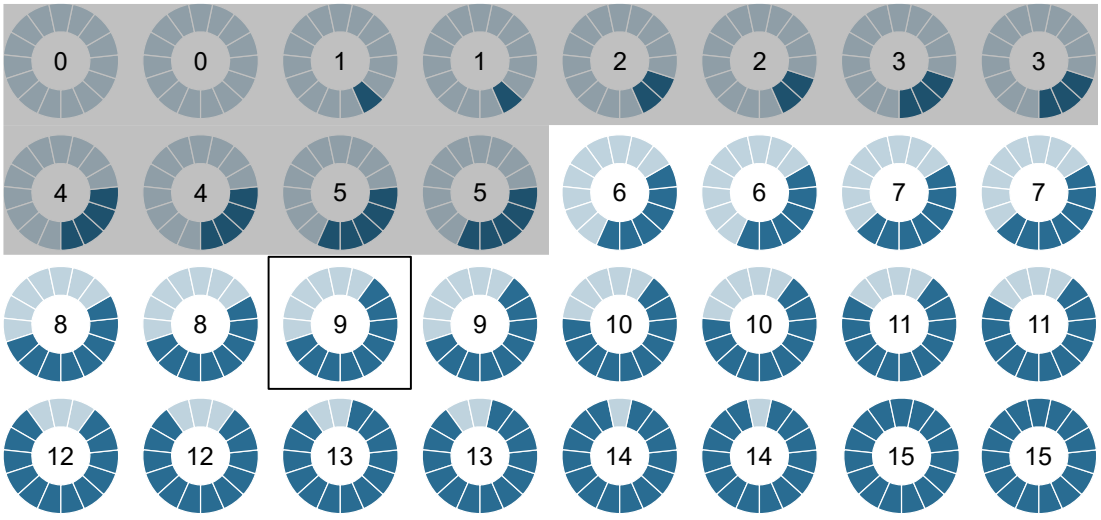
	I disagree a lot	I disagree a little	I neither agree nor disagree	I agree a little	I agree a lot
I would like to explore strange places.	○	○	○	○	○
I like to do frightening things.	○	○	○	○	○
I like new and exciting experiences, even if I have to break the rules.	○	○	○	○	○
I prefer friends who are exciting and unpredictable.	○	○	○	○	○

On the next screen, you will be informed[22]

- which of three situations has been selected, and
- which wheel from that situation has been selected.

Situation 1: the selected wheel was eligible for spinning

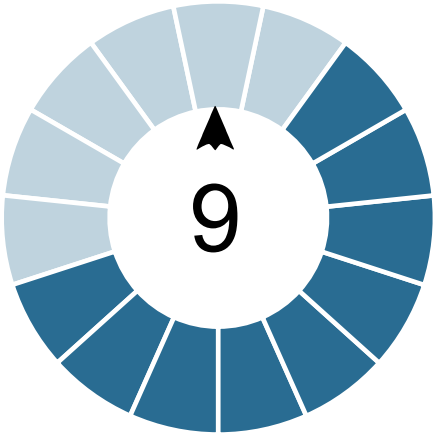
The situation below has been randomly selected. In this situation, you stated that you prefer to spin the wheel if it has **at least 6 dark blue sectors**. The randomly selected wheel is the one surrounded by a black border. Since this wheel is not grayed out (as it has 6 or more dark blue sectors), you will SPIN it for your bonus.



---

[22]Only one of the next two screens was shown to participants, depending on how their randomly chosen decision compared to the randomly selected wheel. The examples illustrate the two possible scenarios.

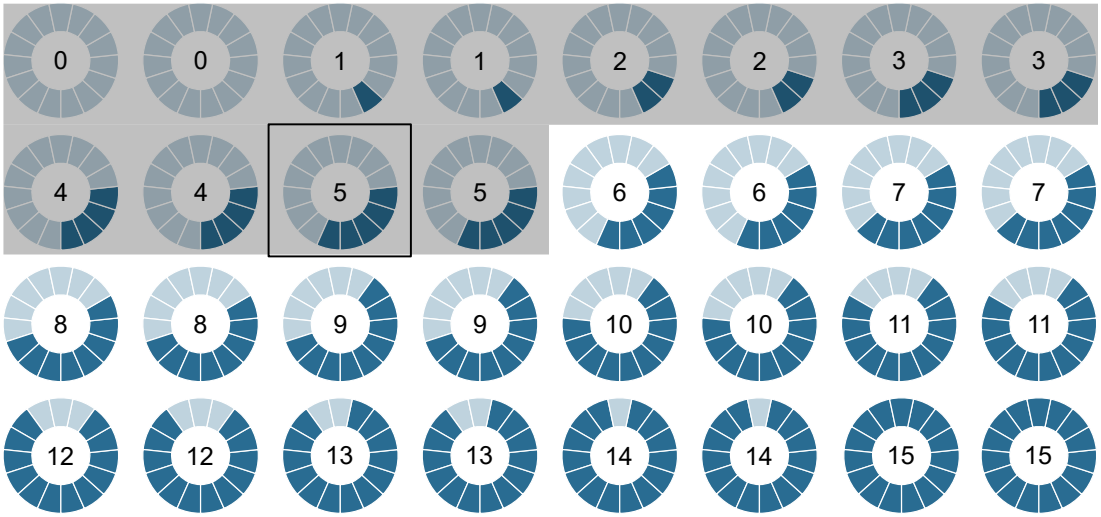
Spin the selected wheel to determine your bonus. If you wish, you can try it out a couple of times before the final spin, which is the one that counts.



Try out Final spin

Situation 2: the selected wheel was not eligible for spinning

The situation below has been randomly selected. In this situation, you stated that you prefer to spin the wheel if it has **at least 6 dark blue sectors**. The randomly selected wheel is the one surrounded by a black border. Since this wheel is grayed out (as it has fewer than 6 dark blue sectors), you DON'T SPIN it. Your bonus is £2.00.



We thank you for your time spent taking this survey. Your response has been recorded.

If you have any comments, please write them in the box below.

If you would like to be informed about the earnings of all participants in this study, please select the option below. We will send you the earnings distribution via Prolific no later than a week after the last submission.

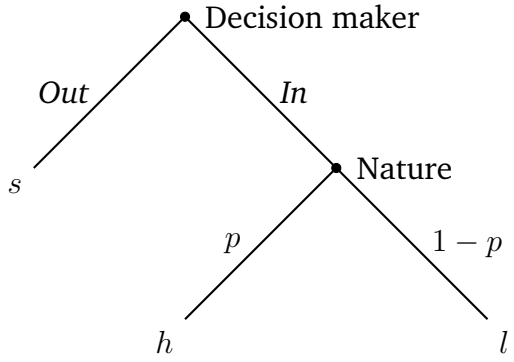
☐ I would like to be informed about the earnings distribution in this study.

4.B. Theoretical benchmarks

4.B.A. The game

We analyze an extension of a simple one-player lottery choice. The decision maker (DM) decides whether to stay *Out* and receive a safe outcome s or to move *In* and play a lottery which pays a high outcome h with probability p or a low outcome l with probability $1-p$. We assume that the DM's utility function $U(.)$ is continuous and differentiable in the set of outcomes. Also, $l < s < h$.

The DM does not know p when making his decision. What he knows is that p is distributed with density $f(p)$ and has full support on the interval $[0,1]$. The DM makes his decision contingent on p. More precisely, we ask him about his minimum acceptable probability, MAP. If p happens to be smaller than MAP, then DM stays Out, otherwise he goes In. The following figure gives a graphical representation.[23]



In the following we look at different benchmarks. In particular, we are interested in whether and how the optimal minimum acceptable probability MAP* depends on the distribution of p.

4.B.B. Expected utility theory

Assume the DM to have a utility function $U(\cdot)$. In an expected utility framework, utility is strictly increasing with outcome. Hence, we have

$$U(l) < U(s) < U(h) \tag{4.1}$$

In this appendix, we consider the MAP as a probability i.e. MAP $\in [0,1]$.

[23] For simplicity, we do not explicitly depict the fact that the DM makes a decision contingent on p.

The DM wants to choose his MAP such that he maximizes his expected utility $U(\mathit{MAP})$ which is

$$U(\mathit{MAP}) = \int_{p=0}^{\mathit{MAP}} f(p) \cdot U(s)\, dp + \int_{p=\mathit{MAP}}^{1} f(p) \cdot [p \cdot U(h) + (1-p) \cdot U(l)]\, dp \tag{4.2}$$

$$= \int_{p=0}^{\mathit{MAP}} f(p) \cdot U(s)\, dp - \int_{p=1}^{\mathit{MAP}} f(p) \cdot [p \cdot U(h) + (1-p) \cdot U(l)]\, dp \tag{4.3}$$

We use the Fundamental Theorem of Calculus to derive the first order condition:[24]

$$\frac{\partial U(\mathit{MAP})}{\partial \mathit{MAP}} = f(\mathit{MAP}) \cdot U(s) - f(\mathit{MAP}) \cdot [\mathit{MAP} \cdot U(h) + (1-\mathit{MAP}) \cdot U(l)] \stackrel{!}{=} 0 \tag{4.4}$$

The density function $f(p)$ has full support. Therefore, $f(\mathit{MAP})$ is positive and we can simplify the expression to

$$\mathit{MAP}^* = \frac{U(s) - U(l)}{U(h) - U(l)} \tag{4.5}$$

The optimal MAP* is independent of the distribution of p. Thus, an expected utility maximizer should not be influenced by it.

4.B.C. Outcome-based add-ons

The result of Section 4.B.B holds if the utility function of the DM is extended by other elements that are based on outcomes. For example, the DM might receive extra (dis-) utility from playing the lottery. Or he might feel additional happiness or regret in case the outcome of the lottery is high or low, respectively. Such add-ons to the utility function lead to a different MAP*, but $f(p)$ would still cancel out of the first order condition. Hence, MAP* should still be independent of the distribution of p.

[24] From the differentiability of $U(.)$ and $\frac{\partial U(\mathit{MAP})}{\partial \mathit{MAP}}(0) > 0$ and $\frac{\partial U(\mathit{MAP})}{\partial \mathit{MAP}}(1) < 0$, we can conclude that at least one local maximum exists. If the solution of the FOC is unique, then it must be this maximum.

4.B.D. Probability weighting

Experimental evidence shows that humans have difficulties in handling probablities. In particular, they seem to overestimate small probabilities and underestimate large ones.[25] In the following we assume the DM to have a continuous probability weighting function $w : [0,1] \to [0,1]$ with $w(0) = 0$ and $w(1) = 1$. This gives the DM the following utility function:[26]

$$U(\textit{MAP}) = \int_{p=0}^{\textit{MAP}} f(p) \cdot U(s)\, dp + \int_{p=\textit{MAP}}^{1} f(p) \cdot [w(p) \cdot U(h) + w(1-p) \cdot U(l)]\, dp \tag{4.6}$$

This leaves us with the FOC:

$$U(s) - [w(\textit{MAP}) \cdot U(h) + w(1 - \textit{MAP}) \cdot U(l)] = 0 \tag{4.7}$$

From the assumptions about the weighting function and $U(l) < U(s) < U(h)$, it follows that this equation has at least one solution (Bolzano's Theorem). As in Section 4.B.B, all solutions are independent of the distribution of p.[27, 28]

4.B.E. Rank-dependent utility

In this section, we present the assumptions we made in order to derive the hypothesis. Unlike in the previous appendices, we focus on a numerical calculation.

We adapt the toy example in Li et al. (2020) and make the following assumptions:

- The utility of outcomes is fixed. We consider $U(high) = U(£4) = 1$, $U(low) = U(£1) = 0$, and $U(safe) = U(£2) = 1/3$.[29]

[25] Back in 2000, Starmer (2000, p. 348–349) mentioned research spanning 50 years which supports this. This still holds true today, e.g. see Li et al. (2020, Figure 4 on p. 276).

[26] Probability weighting is not compatible with the axiomatic framework of expected utility theory. We use the notion of utility in a broad sense covering also non-expected utility theories.

[27] We may get to a unique solution of equation (4.7) if we place additional requirements on the outcomes $U(\cdot)$ and/or on $w(p)$. Two possibilities are: 1.) A unique solution is guaranteed if $w(p)$ is strictly increasing and symmetric, i.e. $w(1-p) = 1 - w(p)$ or 2.) A unique solution is guaranteed if $w(p)$ is strictly increasing and the utility of the low outcome including all possible add-ons is set to zero, i.e. $U(l) = 0$.

[28] In case of multiple solutions, we assume the distribution of p does not offer cues which lead to selecting a different optimum MAP in each treatment.

[29] $U(safe) = 1/3$ was chosen because, given the data on betrayal aversion, it makes a mildly risk-averse player indifferent between accepting any lottery and the safe payoff.

- Participants use a probability weighting function because they perceive the tasks to involve complex risks. Similar to Li et al. (2020), we use Prelec's (1998) *compound invariance* function:

$$w(p) = (exp(-(-ln(p))^{\alpha}))^{\beta}$$

- We use $\alpha = 0.65$ and $\beta = 1.0467$, which according to Li et al. (2020) are the most common values for risky probability weighting.
- Participants use "forward" evaluation: they consider the three possible outcomes and take into account their probabilities, as resulting from the probability weighting function above.
- Participants have the following rank-dependent utility function (Schmeidler, 1989), in which an act generated by a choice of MAP leads to:

$$RDU = w(P(£4)) \cdot 1 + (w(P(£4) + P(£2)) - w(P(£4))) \cdot (1/3)$$

  where $P(£4)$ is the probability of receiving the high payoff for a certain MAP in the respective treatment, $P(£2)$ the probability of receiving the safe payoff, and $P(£1)$ the probability of receiving the low payoff (which does not appear in the utility function, as the utility of the low payoff is considered to be 0).

In this case, the MAPs which maximize participants' utility in the three treatments are: $MAP^*_G = 7$ (RDU = 0.628), $MAP^*_U = 8$ (RDU = 0.495), and $MAP^*_B = 9$ (RDU = 0.439).

4.C. Range-frequency model: a numerical example

According to the range-frequency model, participants categorize stimuli according to range and frequency, and then evaluate them based on a compromise between the two ways of classification.

Let us assume that a participant uses four bins to categorize stimuli: bad lotteries, not OK lotteries, OK lotteries, and good lotteries. When using the range criterion, this participant bins existing lotteries in all treatments in the following way:

Category	Dark blue sectors
Bad	0, 1, 2, 3
Not OK	4, 5, 6, 7
OK	8, 9, 10, 11
Good	12, 13, 14, 15

If she bins lotteries according to frequency, she arrives at the following division in each treatment:

Category	Dark blue sectors
	The Good
Bad	0, 1, 2, 3, 4, 5, 6, 7
Not OK	8, 9, 10, 11, 12
OK	13, 14
Good	15
	The Uniform
Bad	0, 1, 2, 3
Not OK	4, 5, 6, 7
OK	8, 9, 10, 11
Good	12, 13, 14, 15
	The bad
Bad	0
Not OK	1, 2
OK	3, 4, 5, 6, 7
Good	8, 9, 10, 11, 12, 13, 14, 15

If she thinks only categories OK and Good are acceptable and compromises between the two divisions of stimuli, she could report the following MAPs:

$MAP_G = 10.5$, $MAP_U = 8$, $MAP_B = 5.5$. As she gives more weight to the frequency criterion, choices get closer to one another.

Reflections on doing research

While all chapters tackle aspects of how to better cooperate outside our groups, there are two distinct directions in this dissertation: (i) examining the role of exposure to racial diversity in school during adolescence on electoral turnout and on political views later on and (ii) examining the role of betrayal aversion in discrimination in trust, which prompted questions about the measurement of betrayal aversion and about the effects of institutions trying to diminish the scope for betrayal aversion.

I have summarized the findings and their academic and policy implications in the introduction. In this section, I reflect on the lessons I have learned about doing research more generally. I believe these can be useful to junior scientists.

From the studies on betrayal aversion,

□ I have learned the importance of using a correct counterfactual. A natural starting point for experimental economists is to assume participants are rational expected utility maximizers and to work out predictions from there. This is a potentially good enough simplification in some cases, but it might not be in others. When it comes to the control game used to gauge betrayal aversion, evidence indicates it is an oversimplification.

□ Scientists should be modest about what can be inferred from one study, especially when—as is the case in this thesis—findings in one chapter cannot be easily reconciled with those in another chapter. In the experiments in Chapters 2 and 3, I find betrayal aversion at the beginning of an academic year, but not towards its end. The subjects were recruited from largely the same pool and the experiments were similar. Clearly, more research is needed to understand the importance and stability of betrayal aversion.

□ Topic-wise, doing this research has set me on a path of studying how behavioral ethics, psychology and economics interact. In particular, I have thought,

read and discussed about how people draw conclusions about others' intentions from a situation, how they value these intentions, and how outcomes and responsibility allocation (blame/credit) interact to create perceptions of fairness.

From the study of racial diversity in schools,

□ I have been won over by the influential idea that schools' role is not only to create valuable skills for the labor market, but also good citizens (Gradstein and Justman, 2002).

□ The *impressionable years hypothesis*—that there is an optimal window in a person's development when experiences are crucial for further development of certain skills and preferences—is an important takeaway for my future research.

My current research interests combine and further what I have learned from both directions in my doctoral research. Namely, I am designing laboratory (and hopefully later also field) experiments to study spillovers from being exposed to (un)ethical behavior. Ultimately, my plan is to study this in the field, in a school environment. The goal is to evaluate how experiencing ethical behavior in school on the part of the teachers influences active civic behavior later in life, such as the propensity to speak up when observing wrongdoing.

Short Curriculum Vitae

Maria Polipciuc obtained a bachelor's degree in Finance from the Bucharest University of Economic Studies, Romania, and a bachelor's degree in International Economics from the Toulouse 1 University Capitole, France, both in 2008. She studied economics at the master level at Maastricht University and obtained a master's degree in International Economic Studies in 2010. She started working as a part-time teaching assistant at Maastricht University during her studies and continued working full time after graduation until 2013. She then moved to Brussels, Belgium, for a traineeship at the European Commission. She also worked briefly as a research assistant at the University of Liège, Belgium. She moved back to Maastricht to pursue a PhD in economics in 2014. In fall 2018, she visited FAIR (the Centre Experimental Research on Fairness, Inequality and Rationality) at the Norwegian School of Economics in Bergen, Norway. She is currently a postdoctoral researcher at the WU Vienna University of Economics and Business in Austria.

Maria's research focuses on behavioral and experimental economics, discrimination, moral behavior, and hierarchical interactions.

Bibliography

Bibliography

Ackfeld, V. S. (2020). The aversion to monetary incentives for changing behavior. Working Paper No. 100, University of Cologne.

Agresti, A. (2002). *Categorical Data Analysis (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

Aimone, J., Ball, S., and King-Casas, B. (2015). The betrayal aversion elicitation task: An individual level betrayal aversion measure. *PLOS ONE*, 10(9):1–12.

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.

Akbulut-Yuksel, M., Okoye, D., and Yuksel, M. (2020). Social changes in impressionable years and adult political attitudes: Evidence from Jewish expulsions in Nazi Germany. *Economic Inquiry*, 58(1):184–208.

Alesina, A., Baqir, R., and Easterly, W. (1999). Public goods and ethnic divisions. *The Quarterly Journal of Economics*, 114(4):1243–1284.

Alesina, A. and La Ferrara, E. (2000). Participation in heterogeneous communities. *The Quarterly Journal of Economics*, 115(3):847–904.

Alesina, A. and La Ferrara, E. (2005). Ethnic diversity and economic performance. *Journal of Economic Literature*, 43(3):762–800.

Alesina, A., Michalopoulos, S., and Papaioannou, E. (2016). Ethnic inequality. *Journal of Political Economy*, 124(2):428–488.

Algan, Y., Do, Q.-A., Dalvit, N., Chapelain, A. L., and Zenou, Y. (2019). Friendship networks and political opinions: A natural experiment among future French politicians. CEPR Discussion Paper No. 13771.

Algan, Y., Hémet, C., and Laitin, D. D. (2016). The social effects of ethnic diversity at the local level: A natural experiment with exogenous residential allocation. *Journal of Political Economy*, 124(3):696–733.

Allport, G. W. (1954). *The Nature of Prejudice*. Addison-Wesley Pub. Co.

Almås, I., Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2010). Fairness and the development of inequality acceptance. *Science*, 328(5982):1176–1178.

Andreoni, J. and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

Andreoni, J., Payne, A. A., Smith, J., and Karp, D. (2016). Diversity and donations: The effect of religious and ethnic diversity on charitable giving. *Journal of Economic Behavior & Organization*, 128:47–58.

Angrist, J. D. and Lang, K. (2004). Does school integration generate peer effects? Evidence from Boston's Metco program. *American Economic Review*, 94(5):1613–1634.

Armantier, O. and Treich, N. (2016). The rich domain of risk. *Management Science*, 62(7):1954–1969.

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.

Battigalli, P. and Dufwenberg, M. (2022). Belief-dependent motivations and psychological game theory. *Journal of Economic Literature*, 60(3):833–882.

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.

Bellettini, G., Ceroni, C. B., and Monfardini, C. (2020). Immigration, ethnic diversity and voting: The role of individual income. *European Journal of Political Economy*, 61:101840.

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

Bergman, P. (2020). The risks and benefits of school integration for participating students: Evidence from a randomized desegregation program. Working paper.

Bifulco, R., Fletcher, J. M., Oh, S. J., and Ross, S. L. (2014). Do high school peers have persistent effects on college attainment and other life outcomes? *Labour Economics*, 29:83–90.

Bifulco, R., Fletcher, J. M., and Ross, S. L. (2011). The effect of classmate characteristics on post-secondary outcomes: Evidence from the Add Health. *American Economic Journal: Economic Policy*, 3(1):25–53.

Billings, S. B., Chyn, E., and Haggag, K. (2021). The long-run effects of school racial diversity on political identity. *American Economic Review: Insights*, 3(3):267–284.

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

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2013). Under pressure? The effect of peers on outcomes of young adults. *Journal of Labor Economics*, 31(1):119–153.

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.

Boisjoly, J., Duncan, G. J., Kremer, M., Levy, D. M., and Eccles, J. (2006). Empathy or antipathy? The impact of diversity. *American Economic Review*, 96(5):1890–1905.

Bolton, G. E., Feldhaus, C., and Ockenfels, A. (2016). Social interaction promotes risk taking in a stag hunt game. *German Economic Review*, 17(3):409–423.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 90(1):166–193.

Borghans, L., Duckworth, A. L., Heckman, J. J., and ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, 43(4):972–1059.

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.

Brenøe, A. A. and Zölitz, U. (2020). Exposure to more female peers widens the gender gap in STEM participation. *Journal of Labor Economics*, 38(4):1009–1054.

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

Briole, S. (2021). Are girls always good for boys? short and long term effects of school peers' gender. *Economics of Education Review*, 84:102150.

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.

Caeyers, B. and Fafchamps, M. (2020). Exclusion bias in the estimation of peer effects. CEPR Discussion Paper No. 14386.

Cancela, J. and Geys, B. (2016). Explaining voter turnout: A meta-analysis of national and subnational elections. *Electoral Studies*, 42:264–275.

Carrell, S. E. and Hoekstra, M. L. (2010). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics*, 2(1):211–228.

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

Chark, R. and Chew, S. H. (2015). A neuroimaging study of preference for strategic uncertainty. *Journal of Risk and Uncertainty*, 50(3):209–227.

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

Chater, N. and Loewenstein, G. (2022). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, pages 1—-60.

Cheek, N. N. and Norem, J. K. (2017). Holistic thinkers anchor less: Exploring the roles of self-construal and thinking styles in anchoring susceptibility. *Personality and Individual Differences*, 115:174–176.

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

Chuah, S.-H., Hoffmann, R., and Larner, J. (2016). Perceived intentionality in 2×2 experimental games. *Bulletin of Economic Research*, 68(S1):78–84.

Cooper, C. A., Golden, L., and Socha, A. (2012). The big five personality factors and mass politics. *Journal of Applied Social Psychology*, 43(1):68–82.

Corno, L., La Ferrara, E., and Burns, J. (2022). Interaction, stereotypes, and performance: Evidence from South Africa. *American Economic Review*, 112(12):3848–75.

Costa, D. L. and Kahn, M. E. (2003). Civic engagement and community heterogeneity: An economist's perspective. *Perspectives on Politics*, 1(1):103–111.

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.

Cox, J. C. and Sadiraj, V. (2007). On modeling voluntary contributions to public goods. *Public Finance Review*, 35(2):311–332.

Dahlberg, M., Edmark, K., and Lundqvist, H. (2012). Ethnic diversity and preferences for redistribution. *Journal of Political Economy*, 120(1):41–76.

de Rooij, E. A., Green, D. P., and Gerber, A. S. (2009). Field experiments on political behavior and collective action. *Annual Review of Political Science*, 12(1):389–395.

DellaVigna, S., List, J. A., Malmendier, U., and Rao, G. (2016). Voting to tell others. *The Review of Economic Studies*, 84(1):143–181.

Dinesen, P. T. and Sønderskov, K. M. (2015). Ethnic diversity and social trust. *American Sociological Review*, 80(3):550–573.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., and Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550.

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

Dufwenberg, M. and Kirchsteiger, G. (2019). Modelling kindness. *Journal of Economic Behavior & Organization*, 167:228–234.

Easterly, W. and Levine, R. (1997). Africa's growth tragedy: Policies and ethnic divisions. *The Quarterly Journal of Economics*, 112(4):1203–1250.

Edelman, B., Luca, M., and Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22.

Enamorado, T. and Imai, K. (2019). Validating self-reported turnout by linking public opinion surveys with administrative records. *Public Opinion Quarterly*, 83(4):723–748.

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

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maxi-min preferences in simple distribution experiments. *American Economic Review*, 94(4):857–869.

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.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Feld, J. and Zölitz, U. (2017). Understanding peer effects: On the nature, estimation, and channels of peer effects. *Journal of Labor Economics*, 35(2):387–428.

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.

Finseraas, H., Hanson, T., Johnsen, Å. A., Kotsadam, A., and Torsvik, G. (2019). Trust, ethnic diversity, and personal contact: A field experiment. *Journal of Public Economics*, 173:72–84.

Fisman, R., Jakiela, P., and Kariv, S. (2017). Distributional preferences and political behavior. *Journal of Public Economics*, 155:1–10.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.

Friesen, J. and Krauth, B. (2011). Ethnic enclaves in the classroom. *Labour Economics*, 18(5):656–663.

Ge, Y., Knittel, C., MacKenzie, D., and Zoepf, S. (2016). Racial and gender discrimination in transportation network companies. Technical report.

Gerber, A. S., Huber, G. A., Doherty, D., and Dowling, C. M. (2011a). The big five personality traits in the political arena. *Annual Review of Political Science*, 14(1):265–287.

Gerber, A. S., Huber, G. A., Doherty, D., Dowling, C. M., Raso, C., and Ha, S. E. (2011b). Personality traits and participation in political processes. *The Journal of Politics*, 73(3):692–706.

Giuliano, P. and Spilimbergo, A. (2013). Growing up in a recession. *The Review of Economic Studies*, 81(2):787–817.

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

Götte, L., Huffman, D., and Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review*, 96(2):212–216.

Gould, E. D., Lavy, V., and Paserman, M. D. (2009). Does immigration affect the long-term educational outcomes of natives? Quasi-experimental evidence. *The Economic Journal*, 119(540):1243–1269.

Gradstein, M. and Justman, M. (2000). Human capital, social capital, and public schooling. *European Economic Review*, 44(4-6):879–890.

Gradstein, M. and Justman, M. (2002). Education, social cohesion, and economic growth. *American Economic Review*, 92(4):1192–1204.

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. (2004). The role of social capital in financial development. *American Economic Review*, 94(3):526–556.

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.

Guryan, J., Kroft, K., and Notowidigdo, M. J. (2009). Peer effects in the workplace: Evidence from random groupings in professional golf tournaments. *American Economic Journal: Applied Economics*, 1(4):34–68.

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.

Hanushek, E. A., Kain, J. F., and Rivkin, S. G. (2009). New evidence about Brown v. Board of Education: The complex effects of school racial composition on achievement. *Journal of Labor Economics*, 27(3):349–383.

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.

Horowitz, J. K. (2006). The Becker–DeGroot–Marschak mechanism is not necessarily incentive compatible, even for non-random goods. *Economics Letters*, 93(1):6–11.

Hoxby, C. M. (2000a). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, 115(4):1239–1285.

Hoxby, C. M. (2000b). Peer effects in the classroom: Learning from gender and race variation. NBER Working Paper No. 7867.

Jennings, M. K. and Markus, G. B. (1977). The effect of military service on political attitudes: A panel study. *American Political Science Review*, 71(01):131–147.

Jennings, M. K. and Markus, G. B. (1984). Partisan orientations over the long haul: Results from the three-wave political socialization panel study. *American Political Science Review*, 78(04):1000–1018.

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.

Kaplan, E., Spenkuch, J. L., and Tuttle, C. (2019). School desegregation and political preferences: Long-run evidence from Kentucky. Working paper.

Karni, E. and Safra, Z. (1987). "Preference reversal" and the observability of preferences by experimental methods. *Econometrica*, 55(3):675–685.

Kerschbamer, R. and Müller, D. (2020). Social preferences and political attitudes: An online experiment on a large heterogeneous sample. *Journal of Public Economics*, 182:104076.

Kim, Y.-I. and Lee, J. (2014). The long-run impact of a traumatic experience on risk aversion. *Journal of Economic Behavior & Organization*, 108:174–186.

Knack, S. and Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *The Quarterly Journal of Economics*, 112(4):1251–1288.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63(3):190–194.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Kőszegi, B. and Rabin, M. (2007). Reference-dependent risk attitudes. *American Economic Review*, 97(4):1047–1073.

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.

Lavy, V., Paserman, M. D., and Schlosser, A. (2012). Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom. *The Economic Journal*, 122(559):208–237.

Lavy, V. and Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, 3(2):1–33.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622.

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.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review*, 111(6):1807–1844.

Luttmer, E. F. P. (2001). Group loyalty and the taste for redistribution. *Journal of Political Economy*, 109(3):500–528.

Lutz, B. (2011). The end of court-ordered desegregation. *American Economic Journal: Economic Policy*, 2(3):130–68.

Madestam, A. and Yanagizawa-Drott, D. (2012). Shaping of the nation: The effect of Fourth of July on political preferences and behavior in the United States. HKS Faculty Research Working Paper Series 12-034, John F. Kennedy School of Government, Harvard University.

Malmendier, U. (2021). FBBVA lecture 2020 exposure, experience, and expertise: Why personal histories matter in economics. *Journal of the European Economic Association*, 19(6):2857–2894.

Malmendier, U. and Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, 126(1):373–416.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531.

Marschall, M. J. and Stolle, D. (2004). Race and the city: Neighborhood context and the development of generalized trust. *Political Behavior*, 26(2):125–153.

Martinez i Coma, F. and Nai, A. (2017). Ethnic diversity decreases turnout. Comparative evidence from over 650 elections around the world. *Electoral Studies*, 49:75–95.

Marzilli Ericson, K. M. and Fuster, A. (2011). Expectations as endowments: Evidence on reference-dependent preferences from exchange and valuation experiments. *The Quarterly Journal of Economics*, 126(4):1879–1907.

McAuliffe, M. and Triandafyllidou, A., editors (2021). *World Migration Report 2022*. International Organizationfor Migration (IOM), Geneva.

Merlino, L. P., Steinhardt, M. F., and Wren-Lewis, L. (2019). More than just friends? School peers and adult interracial relationships. *Journal of Labor Economics*, 37(3):663–713.

Mondak, J. J. (2010). *Personality and the Foundations of Political Behavior*. Cambridge University Press.

Mondak, J. J., Hibbing, M. V., Canache, D., Seligson, M. A., and Anderson, M. R. (2010). Personality and civic engagement: An integrative framework for the study of trait effects on political behavior. *American Political Science Review*, 104(1):85–110.

Montalvo, J. G. and Reynal-Querol, M. (2005). Ethnic diversity and economic development. *Journal of Development Economics*, 76(2):293–323.

Moreau, M.-C. (2011). What is “language arts”?

Mousa, S. (2020). Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq. *Science*, 369(6505):866–870.

Oberholzer-Gee, F. and Waldfogel, J. (2001). Electoral acceleration: The effect of minority population on minority voter turnout. NBER Working Paper No. 8252.

Oberholzer-Gee, F. and Waldfogel, J. (2005). Strength in numbers: Group size and political mobilization. *The Journal of Law and Economics*, 48(1):73–91.

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

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, 72(6):407–418.

Parducci, A. and Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology*, 89(2):427–452.

Pettigrew, T. F. and Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, 90(5):751–783.

Pettigrew, T. F., Tropp, L. R., Wagner, U., and Christ, O. (2011). Recent advances in intergroup contact theory. *International Journal of Intercultural Relations*, 35(3):271–280.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 66(3):497–527.

Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century. The 2006 Johan Skytte prize lecture. *Scandinavian Political Studies*, 30(2):137–174.

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.

Rao, G. (2019). Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools. *American Economic Review*, 109(3):774–809.

Rogers, T., Ternovski, J., and Yoeli, E. (2016). Potential follow-up increases private contributions to public goods. *Proceedings of the National Academy of Sciences*, 113(19):5218–5220.

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

Scheier, M. F., Carver, C. S., and Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, 67(6):1063–1078.

Schindler, D. and Westcott, M. (2020). Shocking racial attitudes: Black G.I.s in Europe. *The Review of Economic Studies*.

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica*, 57(3):571.

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

Selten, R. (1967). *Beiträge zur experimentellen Wirtschaftsforschung*, chapter Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes, pages 136–168. J.C.B. Mohr (Paul Siebeck), Tübingen, Germany.

Shertzer, A. (2016). Immigrant group size and political mobilization: Evidence from European migration to the United States. *Journal of Public Economics*, 139:1–12.

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.

Steinmayr, A. (2021). Contact versus Exposure: Refugee Presence and Voting for the Far Right. *The Review of Economics and Statistics*, 103(2):310–327.

Stephenson, M. T., Hoyle, R. H., Palmgreen, P., and Slater, M. D. (2003). Brief measures of sensation seeking for screening and large-scale surveys. *Drug and Alcohol Dependence*, 72(3):279–286.

Strahilevitz, M. A. and Loewenstein, G. (1998). The effect of ownership history on the valuation of objects. *Journal of Consumer Research*, 25(3):276–289.

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.

Thomson, K. S. and Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1):99–113.

Tymula, A., Wölbert, E., and Glimcher, P. (2016). Flexible valuations for consumer goods as measured by the Becker–DeGroot–Marschak mechanism. *Journal of Neuroscience, Psychology, and Economics*, 9(2):65–77.

Vigdor, J. L. (2004). Community composition and collective action: Analyzing initial mail response to the 2000 Census. *Review of Economics and Statistics*, 86(1):303–312.

Wenner, L. M. (2015). Expected prices as reference points—Theory and experiments. *European Economic Review*, 75:60–79.

Westfall, P. H. and Young, S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. John Wiley & Sons.

Woods, D. and Servátka, M. (2018). Nice to you, nicer to me: Does self-serving generosity diminish the reciprocal response? *Experimental Economics*, 22(2):506–529.

Yekutieli, D. and Benjamini, Y. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

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

ROA Dissertation Series

1. Lex Borghans (1993), *Educational Choice and Labour Market Information*, Maastricht, Research Centre for Education and the Labour Market.
2. Frank Cörvers (1999), *The Impact of Human Capital on International Competitiveness and Trade Performance of Manufacturing Sectors*, Maastricht, Research Centre for Education and the Labour Market.
3. Ben Kriechel (2003), *Heterogeneity Among Displaced Workers*, Maastricht, Research Centre for Education and the Labour Market.
4. Arnaud Dupuy (2004), *Assignment and Substitution in the Labour Market*, Maastricht, Research Centre for Education and the Labour Market.
5. Wendy Smits (2005), *The Quality of Apprenticeship Training, Conflicting Interests of Firms and Apprentices*, Maastricht, Research Centre for Education and the Labour Market.
6. Judith Semeijn (2005), *Academic Competences and Labour Market Entry: Studies Among Dutch Graduates*, Maastricht, Research Centre for Education and the Labour Market.
7. Jasper van Loo (2005), *Training, Labor Market Outcomes and Self-Management*, Maastricht, Research Centre for Education and the Labour Market.
8. Christoph Meng (2005), *Discipline-Specific or Academic? Acquisition, Role and Value of Higher Education Competencies*, Maastricht, Research Centre for Education and the Labour Market.
9. Andreas Ammermüller (2007), *Institutional Effects in the Production of Education: Evidence from European Schooling Systems*, Maastricht, Research Centre for Education and the Labour Market.
10. Bart Golsteyn (2007), *The Ability to Invest in Human Capital*, Maastricht, Research Centre for Education and the Labour Market.
11. Raymond Montizaan (2010), *Pension Rights, human capital development*

and well-being, Maastricht, Research Centre for Education and the Labour Market.

12. Annemarie Nelen (2012), *Part-Time Employment and Human Capital Development*, Maastricht, Research Centre for Education and the Labour Market.
13. Jan Sauermann (2013), *Human Capital, Incentives, and Performance Outcomes*, Maastricht, Research Centre for Education and the Labour Market
14. Harald Ulrich Pfeifer (2013), *Empirical Investigations of Costs and Benefits of Vocational Education and Training*, Maastricht, Research Centre for Education and the Labour Market.
15. Charlotte Büchner (2013), *Social Background, Educational Attainment and Labor Market Integration: An Exploration of Underlying Processes and Dynamics*, Maastricht, Research Centre for Education and the Labour Market.
16. Martin Humburg (2014), *Skills and the Employability of University Graduates*, Maastricht, Research Centre for Education and the Labour Market.
17. Jan Feld (2014), *Making the Invisible Visible, Essays on Overconfidence, Discrimination and Peer Effects*, Maastricht, Research Centre for Education and the Labour Market.
18. Olga Skriabikova (2014), *Preferences, Institutions, and Economic Outcomes: an Empirical Investigation*, Maastricht, Research Centre for Education and the Labour Market.
19. Gabriele Marconi (2015), *Higher Education in the National and Global Economy*, Maastricht, Research Centre for Education and the Labour Market.
20. Nicolas Salamanca Acosta (2015), *Economic Preferences and Financial Risk-Taking*, Maastricht, Research Centre for Education and the Labour Market.
21. Ahmed Elsayed Mohamed (2015), *Essays on Working Hours*, Maastricht, Research Centre for Education and the Labour Market.
22. Roxanne Amanda Korthals (2015), *Tracking Students in Secondary Education, Consequences for Student Performance and Inequality*, Maastricht, Research Centre for Education and the Labour Market.
23. Maria Zumbuehl (2015), *Economic Preferences and Attitudes: Origins, Behavioral Impact, Stability and Measurement*, Maastricht, Research Centre

for Education and the Labour Market.

24. Anika Jansen (2016), *Firms' incentives to provide apprenticeships—Studies on expected short- and long-term benefits*, Maastricht, Research Centre for Education and the Labour Market.
25. Jos Maarten Arnold Frank Sanders (2016), *Sustaining the employability of the low skilled worker: Development, mobility and work redesign*, Maastricht, Research Centre for Education and the Labour Market.
26. Marion Collewet (2017), *Working hours: preferences, well-being and productivity*, Maastricht, Research Centre for Education and the Labour Market.
27. Tom Stolp (2018), *Sorting in the Labor Market: The Role of Risk Preference and Stress*, Maastricht, Research Centre for Education and the Labour Market.
28. Frauke Meyer (2019), *Individual motives for (re-)distribution*, Maastricht, Research Centre for Education and the Labour Market.
29. Maria Ferreira Sequeda (2019), *Human Capital Development at School and Work*, Maastricht, Research Centre for Education and the Labour Market.
30. Marie-Christine Martha Fregin (2019), *Skill Matching and Outcomes: New Cross-Country Evidence*, Maastricht, Research Centre for Education and the Labour Market.
31. Sanne Johanna Leontien van Wetten (2020), *Human capital and employee entrepreneurship: The role of skills, personality characteristics and the work context*, Maastricht, Research Centre for Education and the Labour Market.
32. Cécile Alice Jeanne Magnée (2020), *Playing the hand you're dealt, The effects of family structure on children's personality and the effects of educational policy on educational outcomes of migrant children*, Maastricht, Research Centre for Education and the Labour Market.
33. Merve Nezihe Özer (2020), *Essays on drivers and long-term impact of migration*, Maastricht, Research Centre for Education and the Labour Market.
34. Inge Ingeborg Henrica Maria Hooijen (2021), *Place attractiveness, A study of the determinants playing a role in residential settlement behaviour*, Maastricht, Research Centre for Education and the Labour Market.

35. Alexandra Marie Catherine de Gendre (2021), *Behavioral Barriers to Success in Education*, Maastricht, Research Centre for Education and the Labour Market.
36. Kim van Broekhoven (2021), *From creativity to innovation: Understanding and improving the evaluation and selection of ideas in educational settings*, Maastricht, Research Centre for Education and the Labour Market.
37. François M.B.M. Molin (2022), *Using digital formative assessments to improve learning in physics education*, Maastricht, Research Centre for Education and the Labour Market.
38. Bart Kasper de Koning (2022), *Empirical Studies on Information, Beliefs, and Choices in Education and Work*, Maastricht, Research Centre for Education and the Labour Market.
39. Alexander Dicks (2023), *NEET in the Netherlands*, Maastricht, Research Centre for Education and the Labour Market.
40. Lynn Lamberdina Johanna van Vugt (2023), *Different NEETs, different needs? Explaining why vulnerable young people are more likely to become NEET*, Maastricht, Research Centre for Education and the Labour Market.
41. Maria-Eugenia Polipciuc (2023), *An exploration of trust, betrayal, & social identity*, Maastricht, Research Centre for Education and the Labour Market.