Sources of Irrational Behavior: Three Essays on Theory and Experimental Evidence

by

Andreas Ludwig

Diploma in Economics, Institute for Advanced Studies (IHS)
Vienna, Austria, 2003
Dipl.-Volkswirt (Diploma in Economics), University of Potsdam
Potsdam,Germany, 2001

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Economics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Spring 2017

Approval

Name:	**Andreas Ludwig**
Degree:	**Doctor of Philosophy (Economics)**
Title:	**Sources of Irrational Behavior: Three Essays on Theory and Experimental Evidence**
Examining Committee:	**Chair:** Dr. Simon Woodcock Associate Professor
	Dr. Jasmina Arifovic Senior Supervisor Professor
	Dr. Jane Friesen Supervisor Professor
	Dr. Arthur Robson Supervisor Professor
	Dr. Erik Kimbrough Internal Examiner Associate Professor
	Dr. Ernan Haruvy External Examiner Professor Naveen Jindal School of Management University of Texas, Dallas
Date Defended:	**April 5, 2017**

Ethics Statement

Professor Steven C. Wright, co-author of the work reported in chapters 2 and 3 of this thesis, has obtained, for the research described in chapters 2 and 3 of this work, human research ethics approval number 38006, granted on December 6, 2007, from the Simon Fraser University Office of Research Ethics.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis.

The original application for the approval and letter of approval are filed with the relevant offices. Inquiries mey be directed to those authorities.

Abstract

This thesis investigates aspects of human behaviour that can be considered irrational from an economic point of view. Potential reasons for three persistent behavioural patterns in economic interactions are investigated: Altruism, discrimination, and punishment of deviant ("immoral") behaviour. For the first two patterns, this thesis reports the results of dictator game experiments with young children in primary schools in Vancouver, BC, Canada. To understand altruism, the thesis looks for potential reasons why children share resources with genetically unrelated others. It shows that socialization in a particular cultural environment, indicated by the language children speak at home, influences children's sharing behaviour to a large extent. The second part investigates discrimination among children belonging to different ethnic groups. It shows that while children from the dominant white category show clear signs of in-goup bias in their sharing decisions, children from the East Asian minority behave based on a more complex ethnic identity. The third part presents a simple game theoretic model to outline a potential evolutionary origin for a genetic disposition to punish behaviour that conflicts with prevailing moral norms. The model shows how human evolution in small groups can make moral punishment evolutionarily advantageous for individual agents.

Keywords: Experimental Economics; Altruism; Discrimination; Child Development; Biological Basis of Economic Behaviour; Moral Punishment

Dedication

This thesis is dedicated to my extended family: My parents Margit and Lothar, who nourished my belief in core liberal values like freedom, justice, and the benefit of hard work in an environment where liberal ideas had a very hard time. My parents allowed me to go my own way, even though it did not look like the most rational path for them at times. My sister Regine, who always supported me, even though I left home so early. My son Jakob, the biggest source of joy and pride I have, and his mother Katrin, who followed me to Vienna and to Vancouver and was a great support over all these years. For you alone I had to finish this work. And finally my lovely wife Isabel, who gave me the energy to set aside the time to get this thesis completed in the end.

Acknowledgements

I am thankful to my co-authors of the empirical work Jasmina Arifovic, Jane Friesen, Steven C. Wright, Lisa Giamo, and Gamze Baray. I also want to thank seminar participants at SFU and at the North American ESA conference 2007, Tucson, AZ, who all provided helpful comments on the early stages of the experimental work. For the work on discrimination, Mohsen Javdani and Benjamin Harris provided the geographic linkage between our experimental data and Census information about neighbourhood characteristics. Brian Krauth and Michele Battisti provided helpful comments on an earlier version of the discrimination work. For the theoretical work, I am thankful to my thesis supervisor Arthur Robson for continued support and many very helpful comments. Financial support was provided by Metropolis British Columbia, the SFU Community Trust Endowment Fund and my former employer, McKinsey&Co.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Economics derives predictions about human behaviour and the dynamics of markets from a very small set of assumptions. Individual maximization decisions under a set of constraints are at the heart of most economic models. Microeconomic models have been applied successfully to explain a wide range of phenomena – from individual purchase and investment decisions, labour market choices to topics adjacent to economics like marriage or voting. There are good reasons to assume that maximization behaviour is at the centre of life itself, as survival of the fittest has been the driver of evolution.

However, not all patterns of human behaviour seem to follow the maximization logic. With the growing use of experimental methods to test basic assumptions on economic behaviour, a significant number of behavioural patterns have been isolated and tested under a variety of conditions that must be considered "irrational" in light of the dominant paradigm in economics. This research focuses on three prominent cases: Altruism in the form of sharing with strangers, discrimination against people purely based on visibly different phenotypes, and punishment of "immoral" behaviour independent of the fact that the infraction of the moral code has no direct impact on the punisher.

Two distinct routes of explanation are investigated – culture and biology. Individuals can be determined to behave "irrational" because they *learn* to do so in a specific cultural context, or because *evolution* has created a strong psychological predisposition. This thesis explores both avenues. Chapters 2 and 3 document empirical work using standard tools of experimental economics. It focuses on differences in behaviour between children with different cultural backgrounds. Chapter 4 models the biological basis of economic behaviour in the specific case of costly punishment of immoral behaviour in a prisoners'-dilemma-like situation. It is inspired by the work of

my supervisor Arthur Robson (for an introduction see Robson 2002). The remainder of this chapter gives a short overview of each of the three main parts of this thesis.

Chapter 2 investigates the behaviour of children in dictator games with hypothetical others. The research is joint work with Jasmina Arifovic, Jane Friesen and Stephen C. Wright and part of an interdisciplinary project on the effects of classroom diversity on children's attitudes and behaviour toward different ethnicities: Jasmina Arifovic, Jane Friesen, Stephen C. Wright: "Education Systems and Outcomes in Diverse Communities: An Interdisciplinary Approach."

In 2007 and 2008, 1,018 children from 100 kindergarten, grade one, and grade two classes in Vancouver, British Columbia, Canada, participated in a dictator game alongside a psychological experiment revealing perceptions of children of different phenotypes. Our experiments took place at the children's schools during their normal school day, allowing us to evaluate the salience and effects of ethnic identities and different experimental treatments on economically relevant behaviour in an important natural setting.

For the results reported in chapter 2, we test the effects of age, gender, language spoken at home, grade, relative height, increased anonymity, as well as for order effects and two other experimental treatments. Our findings indicate that offers in dictator games are influenced the strongest by the language spoken by the participants' parents and the time that each participant was exposed to a public Canadian school environment. Based on these results we argue that altruistic behaviour is, to a significant extent, learned as part of a child's socialization in a specific cultural environment.

Chapter 3 is joined work with Jane Friesen, Jasmina Arifovic, Stephen C. Wright, Lisa Giamo, and Gamze Baray (previously published in Friesen, Arifovic, Wright, Ludwig, Giamo & Baray 2012). The results are derived from a subset of the fieldwork described in chapter 2. We analyze the responses of over 430 Canadian children in a series of activities designed to reveal their evaluations of three ethnic groups (White, East Asian and South Asian), their identification with these groups, and their behaviour towards them in a dictator game. We find that children from the dominant White category have a clear sense of White ethnic identity, and tend to favour White recipients in the dictator game relative to East Asian or South Asian recipients. Minority East Asian children reveal a more complex ethnic identity; they perceive themselves to be equally similar to White and East Asian children. Unlike Whites, East Asian children do not favour recipients from their own East Asian

category, nor do they favour recipients with whom they tend to identify. If anything, East Asian children show out-group favouritism.

Chapter 4 discusses the evolution of morally motivated punishment. It models individuals who are genetically determined to punish particular actions in a prisoner's-dilemma-like situation and shows how punishers of defection could have prospered in the early stages of human evolution, when our ancestors were living in small groups. Their evolutionary success does not only arise from deterring defection and reaping the benefits of cooperation. In the model presented in chapter 4, defection and actual punishment happens; and punishers bear a personal cost. But if punishment is sufficiently efficient, punishers have a *relative* advantage over other players in a small group, even though their net payoff decreases in absolute terms.

The model shows that punishers of defection, who enforce cooperation in bigger groups, eventually fix their types in the entire population. However, a volatile environment is needed for both successful initial mutation and fixation of type to happen with strictly positive probability. If it does, a population of only punishers of defection evolves, which is stable against mutants of any other type. Cooperation will arise as the norm shared by all individuals in the population. The transition to cooperation will be accompanied by permanent hypocrisy - punishers of defection will be the *last* ones to cooperate.

Chapter 5 summarizes the findings documented in this thesis and relates back to the original question on how seemingly "irrational" behaviour can be understood. Are we biologically determined to deviate from individual maximization behaviour or do we have indication that our cultural context shapes our individual actions? Or are both of these forces at work?

The bibliography of this thesis covers the literature used in all three chapters. The appendix provides additional details on the empirical work.

Chapter 2

Socialized to Share. Dictator Games with Young Children

2.1 Introduction

Altruism toward genetically unrelated others is a puzzling feature of human behaviour. Economists are interested in understanding altruism for two main reasons. The first reason is that charitable giving to unrelated others is an important way of allocating resources. Policymakers, as well as recipients of donations, like charities, universities, or hospitals, are interested in the mechanisms that influence people's willingness to share, especially as they see the volume of charitable giving varying between countries. The second reason is that altruism toward unrelated others seems to conflict with maximization behaviour, the standard paradigm in economics for explaining individual decision making. Over the last decades, experimental economists designed a set of simple experiments to explore pro-social behaviour and test for 'pure' altruism in contrast to strategic behaviour aimed at trust and reputation building.[1]

This chapter reports the results of experimental work with children that explores when and where altruism develops. Two field experiments were conducted in 2007 and 2008 with 1,018 children from 100 kindergarten, grade 1, and grade 2 classes in Vancouver, BC, Canada. One experiment was a dictator game and the other experiment was a test of perceptions of children of different phenotypes. The research was part of an interdisciplinary project between economics and social psychology on the effects of classroom diversity on children's attitudes and behaviour toward children of different phenotypes.

[1] See Camerer (2003) for an introduction to the field of experimental economics.

We find that one of the strongest influences on offers in the dictator game is the language spoken by participants' parents at home. The amount of time that a participant spends at a public Canadian school environment has a significant effect as well, even when controlling for the influence of age and measures of integration like the number of classroom friends. Based on these results, we suggest that *socialization* plays a decisive role in the development of altruism in early ages. Socialization is the alignment of attitudes and behaviour with a specific cultural norm over time. For children, this alignment is driven by parents, peers, teachers, and other contacts using different influencing techniques (Hastings, Utendale & Sullivan 2007, Rochat, Dias, Liping, Broesch, Passos-Ferreira, Winning & Berg 2009). Because socialization happens in a specific cultural context, the extent of altruism differs with the home environment of a child. The common influence of attending a general public school system will mitigate these differences over time. The main innovation in this study is our examination of the distributive behaviour of a sufficiently large and rich sample of very young participants to allow for a distinction between the effect of age, just physically getting older, and the impact of the time spent at school as one of the drivers of socialization.

The remainder of the chapter is structured as follows: In section 2.2 we give an overview of the literature related to giving in dictator games, in particular with respect to children as experimental subjects. Section 2.3 contains a description of the sample. In section 2.4 we describe the experimental setup. In section 2.5 we discuss the econometric methodology and present descriptive statistics and the regression results. In section 2.6 we discuss and interpret the results. We summarize our findings in section 2.7. The Appendix documents the scripts that were used in the experiments.

2.2 Literature Review

The gold standard experiment for research on altruism is the dictator game. It is a simple choice of one player over the distribution of an endowment between herself and an anonymous recipient. The first dictator game was conducted by Kahneman, Knetsch & Thaler (1986) as a control experiment to distinguish strategic behaviour in the ultimatum game[2] from 'genuine' altruism. The dictator game was developed into its current, extremely simple form by Forsythe, Horowitz, Savin & Sefton (1994)

[2] In the ultimatum game, like in the dictator game, one player is asked to distribute a certain endowment. However, here the recipient has the option to reject the proposed split of resources, in which case no distribution is implemented and the endowment is lost for both players (Güth, Schmittberger & Schwarze 1982).

and has been used in many different contexts since then. Between 1992 and 2009 alone, more than 120 papers have reported results of dictator games (for a review and a meta-analysis see Engel 2011).

Despite significant reductions in giving due to alterations to the experimental protocol (e.g., Cherry, Frykblom & Shogren 2002, List 2007, Bardsley 2008, Kench & Niman 2009), positive offers in dictator games remain a robust finding in economic experiments. Consequently, the focus changed and today the dictator game is mainly used to explore *why* people share. It is a flexible measuring tool, that can be used to identify cues and triggers for norms that guide pro-social behaviour (Guala & Mittone 2010).

In meta-regressions using aggregated data of 129 studies, Engel (2011) found only a small number of influential variables across all experimental setups. Among those variables that were significant based on more than 20,000 individual observations, age was the strongest predictor of giving in the dictator game. The oldest participants gave significantly more than mid-age participants; students and children gave significantly less.

The possibility that altruistic behaviour might develop with age puts children into the spotlight of social scientists. If altruism is a feature in the ontogeny of human nature, indications of this developmental process should show up in the behaviour of children.

In his seminal work on child psychology, Damon (1977, 1980) conducted interviews with 34 children over a period of three years. He finds that, up until the age of five, participants are mainly self-interested and use arguments of justice only in an ad hoc and self-serving manner. Between the ages of five and seven, Damon finds participants to select strict equality. As children grow older, Damon argues, a more complex notion of justice develops that takes reciprocity and special needs into consideration (for experimental evidence see also Malti, Gummerum, Ongley, Chaparro, Nola & Bae 2016). He maintains that the patterns of moral judgement develop in close interaction with the social world of a child. This basic framework of developmental stages is still guiding most of the research on child development in one form or another (for a brief review of more recent literature, see Murnighan & Saxon 1998, Rochat et al. 2009).

Since the late 1990s, more and more economists and social psychologists have turned to experimental methods to explore pro-social behaviour of children (Gummerum, Hanoch & Keller 2008). Experimental methods in economics reached maturity in testing human behaviour using students as subjects. However, if altruism develops with

age, experimental methods should also be applied to measure altruistic behaviour of children.

Indeed, multiple experimental studies with children find that offers in the dictator game increase with age (Harbaugh, Krause & Liday 2002, Benenson, Pascoe & Radmore 2007, Blake & Rand 2010, Häger 2010), while the variance in the amount offered decreases (Benenson et al. 2007, Harbaugh, Krause & Vesterlund 2007). However, House, Henrich, Brosnan & Silk (2012) test children between the ages of three and eight and find that offers *decrease* with age. The authors argue that children's ability to understand the payoff-maximizing behaviour improves the older they are. For older children (9 to 17 years), Gummerum, Keller, Takezawa & Mata (2008) do not find an effect of age on the amounts offered.

Fehr, Bernhard & Rockenbach (2008) find that children distribute resources very equally at the ages of seven and eight. They attribute this behaviour to an 'inequality aversion', an 'other-regarding preference' that seems to peak in this age group. Rochat et al. (2009) study children in dictator games at the ages of three and five years in seven different countries. They find that considerations for fairness are playing a stronger role for the five-year olds, but are already significant for three-year old children in some cultures. In particular, less than 25% of the younger participants in China, Peru, and Fiji kept all the dictator game endowments for themselves. Pro-social sharing behaviour seems to start developing soon after children have a sense of possession at the age of two (Olson & Spelke 2008, Rochat 2011).

Although most experimental studies agree that the altruism of children increases with age, little is known on what is driving this development. Is it a natural part of human ontogeny, in the form of spontaneously emerging preferences, or rather the result of nurturing?

To the best of our knowledge, there is no direct evidence in experimental economics that the behaviour of children is driven by socialization. However, some papers on the distributive behaviour of children report effects that are consistent with that perspective. Benenson et al. (2007) and Chen, Zhu & Chen (2013) find that children from neighbourhoods with higher socioeconomic status give more when participating in dictator games. Benenson et al. (2007) interpret the differences in offers as a difference in socialization between children living in high and low status environments. Harbaugh et al. (2002) find systematic differences in bargaining behaviour and interpret them as the effect of culture. Also Häger, Oud & Schunk (2012) and Blake, Piovesan, Montinari, Warneken & Gino (2014) document that the acquisition of pro-social behaviour by children is modulated by their socio-cultural

context. Cultural differences in behaviour have also been reported by Roth, Prasnikar, Okuno-Fujiwara & Zamir (1991) who find significant differences in proposals of students playing ultimatum games in different countries.

Additional evidence for differences of child behaviour between cultures can be found in the social psychology literature. Children in more collectivist cultures (China, Fiji, Peru) share significantly more often than children in individualistic societies like the United States or Brazil. (Rochat et al. 2009). In food-sharing experiments, four-year old Chinese children are reported to share more spontaneously (without being requested by the recipient) than Indian children of the same age. Both groups shared more than comparable U.S. children (Rao & Stewart 1999).

All studies discussed in the previous paragraphs find differences that correlate with countries of origin or the socio-economic context of the participants. The authors interpret these variations to be the effects of particular cultures. Our research goes one crucial step further. We use experimental methods to illuminate one of the transmission mechanisms of socialization for young children.

While the main insights derived from our experiments concern the effects of age and socialization, we also test for a number of additional variables that past experimental studies have found to influence the behaviour of young children in a significant way. There is some evidence of gender differences in the behaviour of children in dictator games. Harbaugh et al. (2002), Gummerum, Keller, Takezawa & Mata (2008), Blake & Rand (2010) and House et al. (2012) tested children between the ages of three and 18, and report that girls shared significantly more than boys. Benenson et al. (2007) and Houser & Schunk (2009) find no difference in offers in the standard dictator game between girls and boys. However, Houser & Schunk (2009) find that girls (ages eight to ten) shared significantly more with boys than with other girls, and that girls' preferences for equal sharing are more stable in competitive situations. These results align with observations of adults. There is some evidence for more pro-social behaviour by female participants (Eckel & Grossman 1998, Croson & Gneezy 2009). Croson & Gneezy (2009) also report that women are more sensitive to the context of the experiment and normative clues. Harbaugh et al. (2002) find relative height, measured as the percentage deviation from the mean of their fellow classmates, to be a strong influence for young children. In particular, shorter girls are reported to make dictator offers significantly larger than average.

Not only properties of the dictator matter in experiments with children. Characteristics of the recipient also influence the results. Moore (2009) reports that children, playing Fehr et al.'s (2008) version of a dictator games, shared more with friends than

with acquaintances, and more with acquaintances than with strangers. Chen et al. (2013) confirms this finding. Paulus (2016) even demonstrated that friendship was a stronger incentive to share than need of the recipient. Goeree, McConnell, Mitchell, Tromp & Yariv (2010) conducted dictator games with girls in grades five and six. They find that their sharing behaviour follows a simple 'inverse social distance law'.

The concept of personal identification as a driver of sharing in dictator games has been explored in experimental work with adults (Buchan, Johnson & Croson 2006). Two different hypotheses on the details of the identification effect have been discussed. The first is identification with specific characteristics of the 'other', measured in terms of 'social distance' (Hoffman, Mccabe & Smith 1996). The second is identification with the 'human face' of a recipient that the dictator knows something about (Bohnet & Frey 1999), in contrast to an anonymous 'player B' in a computer lab.

The characteristics of the experimental protocol have been shown to influence giving in dictator games, too. For repeated trials of a dictator game, House et al. (2012) report that sharing decreases with each successive trial. They attribute this result to a learning effect that allows more children to select the payoff-maximizing behaviour. While trial effects could be avoided by playing dictator games only once, most experimental economists maintain that repetition is one of the cornerstones of sound experimental conduct (Hertwig & Ortmann 2001).

The dictator game in our study was combined with a second experiment in a longer testing session. Consequently, we needed to control for potential order effects between the two experiments. Half of our participants played the dictator game last. The other half, however, started with the dictator game and knew that it would be followed by a second task that would only be explained after completing the dictator game (the 'sharing task'). Although the level of sharing had no consequences for the participants other than determining their own final payoff, fear of retaliation might still explain positive offers (Fehr & Gachter 2000, Dahlman, Ljungqvist & Johannesson 2007, Smith 2007). To control for (expected) reciprocity, decisions in dictator games are usually made anonymously. Some researchers maintain that anonymity is a fundamental prerequisite to accurately measure the level of altruism (e.g., Eckel & Grossman 1998).

Finally, how an experiment is described can potentially prime participants and influence behaviour. 'Psychological prominent' focal points were introduced into the game theory literature by Schelling (1960) and became well known as 'framing' in psychology with Tversky & Kahneman's paper "The framing of decisions and the psychology of choice" in 1981. Today it is common knowledge among both psycholo-

gists and behavioural economists that the framing of a situation can strongly influence decision making. In more recent work, List (2007) and Kench & Niman (2009) show that offers in dictator games can even become negative by changing the frame of reference and allowing the participant to give to, but also to take from the recipient.

30 years since it was played for the first time, the dictator game became a standard tool for experimental economists. This research will contribute to the large body of literature with what we consider the most comprehensive application of a dictator game to the study of child behaviour. To the best of our knowledge, our study surpasses previous experimental work on altruism among children both in size of the experimental sample and in the number of behavioural hypotheses that are tested simultaneously.

2.3 Sample Description

2.3.1 Participants

Participants for our experiments were recruited from 100 kindergarten to grade 2 classrooms in 38 public elementary schools in Vancouver, British Columbia, Canada. These 38 schools represent 42% of all public elementary schools and annexes in the Vancouver School District. Permission was granted by the Vancouver School District to recruit participants conditional on the agreement of schools, teachers and parents. Schools were chosen based on location in oder to cover all major neighbourhoods in Vancouver. As a result, the sample captures the variety in Vancouver of children's phenotypes, home language, and socio-economic characteristics.

Parental informed consent was obtained prior to entering any classroom. In 83 of the 100 classrooms, all children were invited to take part in the study; 18 children from 17 classrooms were recruited on an individual basis.[3] In classrooms in which all children were invited to participate, 72% of parents gave written consent for their children to participate. Some of the children could not be tested because they were absent on the day of testing, so that the overall participation rate was 70%.[4] The experiments were run over the course of two years: April 17 to June 19, 2007 and January 9 to June 26, 2008. A total of 1,088 children were tested, 167 in 2007, 893

[3] This was done to facilitate a related project by increasing the number of children who took part in a lottery for one of the Vancouver school board's magnet programs, but did not get a spot in their first preference school.

[4] Response and participation rates are calculated based on the 75 classrooms with reliable information on total class size. For eight classrooms, this part of the teacher interview information was not available.

in 2008, and 28 both in 2007 and 2008. In 1,043 testing sessions, the dictator game was completed.

Selected demographic characteristics of all children completing the dictator game are reported in Table 2.1 on page 36. There were slightly more boys in the sample than girls. Over 60% of the participating children were of East Asian or Caucasian decent. The third main phenotypic group, children of South Asian ethnic background, only consisted of 74 participants (7.1%).

For less than half the participants, English was the only language spoken at home. Table 2.2 provides a detailed break-down of English as exclusive home language by phenotype. According to their respective teachers, 83% of all participants were fluent in English, 14% had a working knowledge and only 3% had problems communicating in English with their class mates. If testers noticed language problems preventing a child from properly understanding a task, individual result were flagged and excluded from the analyses.

The average age of all participants was 6.1 years. Table 2.3 provides a detailed overview of the age distribution by phenotype, grade, and home language. Table 2.4 shows the results of a Kolmogorov-Smirnov equality-of-distribution test for the age distribution of participants from English-speaking and Non-English-speaking families. Looking at the entire sample of participants, no significant difference in age distribution can be detected. However, as the t test reported in table 2.5a shows, participants from non-English-speaking families are significantly older in kindergarten. The two older cohorts in grade 1 and grade 2 do not show this significant difference in age.

Tables 2.6 and 2.7 present selected results of the sorting task. These results characterize the level of social integration of the participants. While Table 2.6 shows the distribution in the number of classroom friends of Caucasian, East Asian or South Asian phenotype, Table 2.7 shows how many times participants considered one of these dominant phenotypes as 'like you.' Both indicators are used to measure social integration of our participants. While the number of classroom friends measures the observable integration into the classroom environment, the number of pictures of hypothetical others from the three major phenotypes that have been selected as 'like me' indicates the child's perceived integration with the mainstream culture of Vancouver. The number of participants with no or only one classroom friend is clearly higher among children from non-English-speaking families. The measure of perceived integration shows no obvious difference between the major ethnicities.

2.3.2 Testers

A total of 22 experimenters took part in the field work; 14 females and 8 males. They were recruited from the pool of graduate and undergraduate students in the departments of Psychology and Economics at Simon Fraser University. Fourteen testers were of Caucasian decent, two were of East Asian, two of South Asian, one of Hispanic/Latin American, two of Middle Eastern, and one of mixed Caucasian/East Asian decent.

All testers took part in a training process to ensure that everyone was following the scripts as closely as possible and all testers were comfortable working with the age group of interest. Teams of three to four experimenters (mixed gender, department, and phenotype as far as possible) were established on a school-by-school basis.

2.4 Experimental Design

2.4.1 General Design

The dictator games were part of an interdisciplinary research project on the influence of classroom diversity, inter-ethnic contact, and friendship on the attitudes and behaviour of children toward other children with visibly different phenotypes (see also Giamo & Wright 2008, Friesen et al. 2012). The field work consisted of three main parts: A sorting task, the dictator game, and accompanying teacher interviews.

In the sorting task, children participated in a survey-like measurement of their attitudes. Each child was presented a set of 12 pictures of children his/her age with clearly Caucasian, East or South Asian phenotype. Then the participant was asked to indicate which of the presented children he or she thought had certain competencies and characteristics, like 'smart', 'bad', 'has lots of friends', 'helpful', etc. For each question, a tester shuffled the pictures and presented them in a random arrangement on the table in front of the participant. Subjects were also asked if they are worried about the children in the pictures, how similar to themselves they perceive them to be, and whether they would like any of the children as a close friend. Before the experiment, the tester asked the participant to indicate all girls and all boys to ensure that the participant understood the concept of the sorting task. A total of 16 properties were tested in approximately 20 minutes.

In the same session, either before or after the sorting task, the participant played a dictator game against hypothetical others represented by individual pictures of children with clearly Caucasian, East or South Asian phenotype and same sex as the

participant. Their pictures were printed on envelopes. Participants were endowed with stickers and asked if they want to share with the hypothetical others. If a participant wanted to share with any of the hypothetical others, they were asked to put the stickers into the respective envelops and the envelops into a box collecting all stickers shared with hypothetical others. The dictator game was repeated over three or six trials, under different treatments, which we discuss in further detail below.

In parallel to the experiments, teachers were asked to provide information about the characteristics of each participant, such as gender, phenotype, home language and English proficiency, friendship patterns, or relative height.

The tests were conducted during normal school hours, between nine am and three pm. The teacher informed the participants in each class about the procedure and introduced the testing team. Each child was escorted from his/her classroom by a tester and led to an empty classroom, the school library or the hallway. Before beginning the experiments, the tester introduced her/himself in a standardized way, asked the child's name and age, and took a picture of the participant. The participant's photograph was printed and added to the picture package. It was used in the attitudes part of the sorting task to measure the subject's self perception. Participants were allowed to keep their pictures.

Both experiments were introduced to the participants as a series of games. Ninety-two percent of the participants in the dictator game were also asked to do the sorting task. The other 83 children played a reciprocity game in the second part of the testing session.

Testing was done in a face-to-face manner with the tester and the child both sitting at a table together. The sessions lasted no longer than 30 minutes. The sorting task took 15 to 20 minutes, while the sharing task took between 5 and 7 minutes. In pre-tests of the procedures, 30 minutes proved to be the maximum attention span of an average child age 5 to 7, which restricted the number of possible repetitions and treatments we could introduce. In the rare case that a child was not able to follow the procedures, the tester gave some stickers to the child, as well as their picture and returned the child to the classroom.

2.4.2 Pictures To Represent Hypothetical Others

A total of 18 different photographs were used in our experiments; 12 in the sorting task and two times three in the dictator game. In the sorting task, participants saw two boys and two girls of each phenotype – Caucasian, South-Asian and East-Asian.

In the dictator game, a participant only saw three hypothetical others, all of the same gender as her/himself, one of each phenotype.

All pictures had been rated by graduate students of Simon Fraser University according to phenotype, perceived gender, age of each child, and, on a five-point Likert scale, attractiveness, and facial expression. The goal was to generate a set of pictures with maximum perceived homogeneity in all dimensions except for phenotype. Nine student testers (ages 20 to 30) rated a total of about 350 different pictures from various stock photography websites. Three of the students had a Caucasian ethnic background, three East Asian, two South Asian, and one was of African American decent.

If one of the testers was uncertain if the child in a picture was between age five and seven, or as to the sex of the child or their phenotype, the picture was excluded from further consideration. Out of the remaining pictures, we selected five sets. Each set contained six pictures, one of each gender and phenotype. All of the pictures in a set had a Likert score less than one point away from the set average in both attractiveness and facial expression. Two sets of pictures were randomly selected for the sorting task while one set was randomly selected for the dictator game.

2.4.3 Dictator Game

This section introduces the dictator game in more detail. It describes how we adapted the experiment for our work with children. We also provide an explanation of different treatments that we introduced to test potential reasons to share. The treatments were composed of a combination of four variations to the testing protocol. Variation one changed the way that hypothetical others were presented to the participants. Variation two changed which trial of the experiment was conducted under increased anonymity. Variation three determined if additional information was given on the hypothetical others. The forth variation changed the order in which sorting and sharing task (the dictator game) were conducted. Combining the individual variations, we used a total of 13 different treatments in our dictator games. Each treatment was detailed in an experimental script.

Guided by the experimental scripts, each participant was tested individually by one of our testers. For a given treatment, the dictator game was always introduced to the participants with the same words. The scripts required the testers to be verbatim at this point of the experiments to frame the experiment in the same way for each participant.

The participants played the dictator game against hypothetical others. Colour pictures of the hypothetical others were printed onto envelopes to facilitate anonymous trials in the experiment. The tester presented three photographs of other children with the same gender as the participant, but different phenotypes (Figure 2.1). The pictures that were used in the dictator game were different from those used in the sorting task.

The participant received a new endowment of stickers in each trial of the dictator game; 36 stickers in total over all trials. We chose stickers as an incentive for practical reasons. Any kind of food was ruled out because of allergy concerns. Money was not used because it was impossible to ensure parental control over how the children would spend the money (see Benenson et al. (2007)). Based on feedback from pre-tests in a class of children in a neighbouring school district, a set of stickers was selected that proved to be attractive to participants of our age group (Figure 2.2).

The tester explained to each participant that she/he could keep the complete endowment of stickers or share them with any or all of the hypothetical others. Participants were informed that the stickers they decided to share were distributed to children looking like the ones in the pictures. The participants also learned from the introduction whether the sharing task was the first or the last activity in the session.

The tester was not allowed to interrupt the child by asking if he or she was finished distributing the stickers, but needed to wait for the child to indicate it. This was done to prevent the child from interpreting the question as an indicator of tester approval of a given distribution of stickers.

Variation 1: Simultaneous versus sequential sharing

For most participants (63.6%), the pictures of the three hypothetical others were shown simultaneously. The participants were asked to distribute twelve stickers among the three hypothetical others and themselves. We chose to use twelve stickers to allow for different sharing patterns. Equal distributions between two, three or four individuals were possible. At the same time the endowment was large enough to allow for variability in unequal distributions. The procedure was repeated three times. The testers randomly changed the arrangement of the pictures on the table for every trial. We call this treatment *simultaneous sharing*.

In the *sequential sharing* treatment, participants were asked to distribute an endowment of six stickers between one hypothetical other and him/herself. The order of presentation of the hypothetical others was determined on the result sheet that

the tester was given before testing the next child. The task was repeated six times, so that the total endowment over all six trials was again 36 stickers.

Variation 2: Perceived anonymity

One of either trial two or three in the simultaneous treatment and three randomly selected trials in the sequential treatment were conducted under increased anonymity. In the respective trial, the child was not directly observed by the tester, who turned round and did not watch the participant perform distributing the stickers. The participant put the stickers s/he wanted to share into the envelope of the respective hypothetical other and the stickers s/he wanted to keep into a blank envelope. Then the participant put the envelope(s) for the hypothetical other(s) into a box with a large number of other envelopes, including the ones from previous trials.

The tester explained the increased anonymity to the participants. The tester reassured the participant that parents, friends and teacher would never know her/his choices.

Variation 3: Additional information on hypothetical others

In 479 cases (45.9%), the participants in the dictator game were read out additional information about the hypothetical others. This information consisted of a name for each hypothetical other,[5] and favourite sport or activity (Table 2.8). The additional information were randomly matched to the hypothetical others. The tester read it to the participant the first time that the respective hypothetical other was shown. The tester was allowed to repeat the information at the participant's request.

At the end of a session with a participant who was exposed to the additional information, the tester conducted a short exit interview to learn if the participant knew children with any of the names used for the hypothetical others, or shared any of the preferences for favourite ice creams or activities.

Variation 4: Control for order effects

Fifty-five percent of all participants started with the dictator game, while the other forty-five percent started with the sorting task.

[5] In order to select names that were as neutral as possible, the main author conducted an online survey where ethnically neutral names could be suggested in a first stage. The names suggested in stage 1 were rated in a second stage of the survey to select names that would fit to children of Caucasian, East Asian, and South Asian decent equally well. Forty-two people from the Greater Vancouver area participated in the first stage of the survey, while 49 participated in the second.

We created 13 scripts with different treatments and testing orders and assigned them randomly to individual participants and testers to control for the possible effects of the presentation order of the two tasks. All four variations were randomized in our scripts: (1) simultaneous or sequential sharing, (2) the trial(s) in which the dictator game was run under increased anonymity, (3) the use of additional information about the hypothetical others, and (4) the order of the sorting task versus the sharing task.

With a total sample of more than 1,000 participants we could obtain a sufficient number of observations for each of the 13 scripts and were able to test for the effects of each variation in a controlled manner. The next section will discuss the results in detail.

2.5 Results

2.5.1 Descriptive Statistics

Figure 2.3a presents the main summary statistics and a histogram of the total number of stickers that the participants decided to give away. The modal responses across all participants were to give away 9 (25%) or 18 (50%) out of 36 stickers; the mean was 13.8 stickers shared (38.4%). In over 36% of the cases, the participants shared at least half of their stickers; only 5.2% of the participant kept all the stickers over all trials.

Figures 2.3b and c show the number of stickers shared in each trial. Under simultaneous sharing (Figure 2.3b) the frequency distribution has peeks at multiples of 3. For sequential sharing (Figure 2.3c), the modal response is 3, which is 50% of the endowment per trial.

2.5.2 Regressions

The dependent variable in all of our regressions is the number of stickers shared. Because the total endowment of stickers over all trials in both the simultaneous and the sequential treatments was the same, the data were pooled. Most of the analyses are based on the number of stickers that each participant shared out of a total of 36. We call these tests *student-level* regressions. Three additional datasets are used to tests for trial- or hypothetical-other-specific effects. The datasets are differentiated by the aggregation level of the dependent variable *stickers shared*. Table 2.9 provides an overview. We refer to the respective statistical analyses as *trial level* and *trial and picture* level regressions.

In all cases, we are dealing with count data. The dependent variable is discrete and non-negative. The most appropriate regression models for this type of data are of the Poisson or negative binomial type (Long & Freese 2005). In this research, we use a hurdle model consisting of a logit regression and a zero-truncated negative binomial (ZNTB) regression. The logit model regresses possible explanatory variables on the probability to share nothing at all. The ZTNB part regresses the number of stickers shared, conditional on a positive offer, on possible explanatory variables. There are two reasons for using a hurdle model:

1. For all the specifications tested in this research, the data exhibits significant over-dispersion.[6] Over-dispersion can be accounted for by using either a hurdle model or a zero-inflated Poisson/negative binomial model. Zero-inflated models are inappropriate, however, as they assume the existence of a type of participant who never ever shares.[7] Contrary to this hypothesis, we assume that the behaviour of each child is influenced in the same way and both sharing zero and sharing a positive amount is a possible outcome of the experiment for any given participant.

2. The tester asked the participant what he/she wanted to do after explaining the procedure. Most children answered either 'I take all the stickers,' or 'I want to share.' Only if they declared that they wanted to share would the participants start distributing sticker. The hurdle model accounts for the possibility of a two-step decision-making process. First the participant decided if he/she wanted to share at all, and only then decided about the amount to be shared. (Compare Engel (2011) and Blake & Rand (2010) who derive similar conclusions about the decision-making process in dictator games and use similar estimation methods.)

[6] The assumtion required for using a simple Poisson model is that the variance equals the mean. This is violated in our experimental results. For student-level data, the variance of 65.4 is clearly larger than the mean of 13.8. Formally we tested for over-dispersion by fitting a (non-truncated) negative binomial model and a comparison Poisson model in each specification and used a likelihood ratio test to see if the over-dispersion parameter α in the ZTNB model is significantly different from zero. The hypothesis $\alpha = 0$ was always rejected at the 99% confidence level.

[7] Zero-inflated models assume that the excessive amount of observations of zeros are the result of sampling from a pool of two types of participants; one type never sharing and the other sharing from 0 to 36. We assume, however, that a type of children who strictly share nothing, irrespective of any external influences, does not exist. In fact, 20.3% of the participants decided to share nothing in some, but not all the trials of the experiment.

Estimation of the hurdle model is done in two steps. First, a logit model is estimated using the maximum likelihood method, fitting an equation of the form:

$$logit(\pi_i) = \ln \Omega(\pi_i) = \ln \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x_i}\beta + \varepsilon_\mathbf{i}, \tag{2.1}$$

with $\Omega(\pi_i)$ being the odds ratio of sharing any sticker per participant (student-level regression), trial or trial and picture of hypothetical other, respectively.

The results reported in Tables 2.10, 2.12, 2.14a, 2.15a, and 2.16 are the individual elements β_k of β, the raw coefficients estimated for the logit models. All the marginal effects for the logit models are given as changes in the *logit* $\Omega(\pi_i)$, not in probabilities. The changes in Ω for a unit change in the independent variable do not depend on the levels of the other variables in the model, while the changes to the probabilities do (Long & Freese 2005, p. 177ff). As the interpretation of the raw coefficients is not the most intuitive, we report percentage changes in the odds ratio with unit changes to the explanatory variables for significant effects.

Second, a zero-truncated model is estimated using maximum likelihood, based on only the cases of non-zero values for stickers shared. The ZTNB model has the functional form:

$$\tilde{\mu_i} = e^{\mathbf{x_i}\beta + \varepsilon_\mathbf{i}}, \tag{2.2}$$

with the results reported in Tables 2.11, 2.13, 2.14b, 2.15b, and 2.17 being the raw coefficients β_k again. Also in this case, we report percentage changes to the expected number of stickers shared for a unit change in the explanatory variable for selected cases. All marginal effects in the non-linear models have been calculated using the STATA tool listcoef provided by Long & Freese (2005).

The impact of socialization is identified in this research by contrasting age, exposure to school environment, and family background of the participant. Age was reported by each participant before the experiment. The low ages of the participants allows us to distinguish two drivers of socialization: a child's family background and the public school environment.

As a proxy for differences in family values, we use the participant's home language. All major[8] languages other than English are used in specification 0 in Tables 2.10 and 2.11. In specifications 1a to c and all subsequent regressions, we include a binary variable indicating at least one language other than English is spoken at home (value

[8] Languages spoken by more than 20 participants (or 2% of the entire sample).

1) or exclusively English as the home language (0). English as the exclusive home language is also a proxy for a family's immigration status. While a language other than English is spoken in most families of first generation immigrants, it is less likely to be preserved among second or later generations (Houle 2011).

The effect of age is distinguished between participants from exclusively English-speaking families and families with other home languages. If non-English-speaking families teach different values with regard to sharing than parents only speaking English, longer exposure to these family values should widen the gap to mainstream behaviour.

Exposure to public school education is measured as the participant's grade.[9] If the values taught in exclusively English-speaking families are different from those taught in families in which (one of) the language(s) spoken is different from English, then the effect of exposure to the school environment might vary with the home language of a participant, too. Therefore, the variable *grade* is interacted with the variable *non-English home language*.

A subtle possibility for an unobserved variable impacting both grade conditional on age and sharing behaviour is related to what is known in the educational literature as 'red-shirting' (Lincove & Painter 2006, Deming & Dynarski 2008, Kluczniok 2012). If families that are less assimilated to Canadian mainstream culture are more likely to defer school entry for their children, then participants from less assimilated families are generally in lower grades then their same age peers from mainstream culture families.

If children from less assimilated families systematically differ in their sharing behaviour, then the effect of socialization in their school environment might be overstated. A child from a more assimilated family might have entered school earlier, and is already in grade 2, while a child of the same age from a less assimilated family has been red-shirted and is still in grade 1. If a child shares more in grade 2 then the same-age participant of the same phenotype in grade 1, the effect could be attributed to exposure to school environment, while in fact it is the impact of a different family background.

Unfortunately, our information about the participants' ages is not precise enough to identify children who were older than the required age when they entered school. Also, we use grade to measure exposure to the Canadian public school system, al-

[9] Using grade as a meassure of exposure to the Canadian schooling system is only a proxy, as some of the children might have entered a Canadian primary school only in grade 1 or 2. Children might also have repeated a grade.

though some children might have repeated a class while others might have immigrated and only joined grade 1 or 2. Thus, it is not possible for us to control for red-shirting in our regressions. However, as the test reported in Table 2.4 shows, there is no significant difference in age distribution between the two groups of participants: participants from families with non-English and participants from families with English home language. Children from non-English-speaking families are significantly older only for the cohort of participants from kindergarten classes (Table 2.5a). For grade 1, and grade 2, there is no significant age difference (Tables 2.5b and 2.5c). If a significant number of participants from less assimilated families entered school late or repeated classes, because they were held back by their parents, the entire age distribution in the first group would be significantly shifted upwards compared to the second group.

Specification 1c of the logit and ZTNB regressions reported in Tables 2.10 and 2.11 explores additional socialization effects in more detail. Differences in behaviour between children coming from families with different home languages can be interpreted in at least two ways. On the one hand, differences in behaviour could be driven by different systems of values. These values are part of the cultural context of a child, which may be indicated by the individual languages spoken in a child's family. Over time, exposure to a common educational environment might reduce the differences in values that children hold and that guide their behaviour. On the other hand, there might be differences in behaviour that have nothing to do with different values. All children might share more with individuals they consider to be part of their respective 'in-groups' (Harrod 1983, Giamo & Wright 2008). This might be a universal feature across all cultures (but see also Spielman 2000). Immigrant children might change their behaviour as soon as they feel more comfortable in their new environment. They might start sharing more when they stop being an outsider.

To distinguish the two potential channels of socialization, we include two more variables. First, the number of classroom friends with the same phenotype as the hypothetical others who the participant is asked to share with. This variable measures *observed* integration of children into the mainstream Canadian culture. The second variable is the number of times one of the pictures with children of Caucasian, East or South Asian phenotype is selected as 'children like you' in the sorting task. This variable measures the perceived similarity of a particular participant with children which belong to one of the three dominant phenotypes in the lower mainland of British Columbia. It is a proxy for *subjective* identification with the main stream.

All other variables of interest are tested as main effects in specifications 2 and 3 of the student-level regression (Tables 2.12 and 2.13). Trial- and hypothetical other-specific effects are explored separately (Tables 2.14, 2.15, 2.16, and 2.17).

Gender information was recorded by the tester and verified during the teacher interviews. A participant's height was assessed *relative* to her/his classmates by the participant's teacher.[10] Thus, in the same way as the variable used by Harbaugh et al. (2002), the variable used here is *not* height relative to a recipient in the dictator game. If there is any effect of relative height, it is not situational in the game played, but a persistent effect of a classroom environment on a child's behaviour, similar to the impact of socialization.

Treatment differences are indicated using dichotomous dummy variables which are listed in detail below. Of course we can only test *joint* hypotheses in our regressions. To gather evidence that participants shared for a particular reason, we need to establish that an indicator for an experimental treatment has a statistically significant effect on the amount of stickers shared *and* be convinced that the indicated *treatment is effective* in measuring the underlying real-world condition of interest. We will come back to this problem when interpreting the results.

Indicators of testing order and the use of additional information to describe hypothetical others are included in specification 3 of the student-level regression and all specifications of trial-level and trial and picture-level regressions. A dummy variable for sequential sharing is included in all the regressions except for the trial-level regressions as they are run separately for simultaneous and sequential sharing results. Increased anonymity is only considered in trial-level regressions (Tables 2.14 and 2.15).

Identification with the hypothetical other is included in regressions using the most disaggregated data, reported in Tables 2.16 and 2.17. Specifications 1 and 2 of these student and picture-level regressions test the two prominent hypotheses on how identification with a hypothetical other could lead to increased sharing: the 'human face' hypothesis (Bohnet & Frey 1999) and the concept of 'social distance' to specific properties of the hypothetical others (Hoffman et al. 1996).

We test the first hypothesis by examining if the inclusion of any additional information on the hypothetical others has a non-negative effect on offers in the dictator game. The effect does not need to be positive, as the participant could already be

[10] For logistical reasons we did not measure the height of each child directly. Therefore, we cannot replicate the variable used by Harbaugh et al. (2002) exactly; we use the teacher information as a proxy.

convinced that the hypothetical other is a real person. But it must not be negative under the hypothesis that dictators share the more the less abstract the recipient of their sharing becomes.

To test for the impact of social distance, however, characteristics or preferences need to be identified that are shared between hypothetical other and participant. For this purpose, we use information gathered in exit interviews with every participant who received additional information on the hypothetical others. In particular, social distance could reduce because the participant knows somebody with the name given to a hypothetical other, or because participant and hypothetical other share preferences for a particular kind of ice cream, sports or activity.

In addition to the possible reduction of social distance due to the additional information on the hypothetical others, we examine whether phenotype is a dimension of social distance that is relevant for sharing behaviour. For this purpose, we distinguish an external classification of participant and hypothetical other as belonging to the same phenotype from the participant's own assessment of similarity to the hypothetical other. The first is based on teacher information on the participant's phenotype, the latter on the fact that the participant identified children of the same phenotype as the hypothetical other as 'like me' in the sorting task.

The only critical influences on human behaviour in experiments that we described earlier but cannot control for are framing effects. The language used to introduce the participants to their tasks had been fixed in the scripts to ensure comparability of results between testers. The phrases were calibrated in pre-tests to make the participants share at least some of their stickers with the hypothetical others.[11] Using this methodology, we were able to measure differences in sharing due to different treatments and demographic characteristics, even though framing influences the absolute amount of stickers shared.

As an additional measure to take possible framing effects into account, we checked all results for robustness against tester effects. The difference between testers could be a source of bias due to a slightly different explanation of the experimental procedures, or even just differences in tone of voice and assertiveness in behaviour. For the sake of brevity, the results of the robustness checks are not reported here, but are available on request.

[11] The most influential framing in the procedures seemed to be the prominent use of the word 'share' instead of more neutral terminology like distribute or give.

2.5.3 Regression results

We first turn our attention to the main variables of interest. The effects of socialization on the total amount of stickers shared can be distinguished in the logit and ZTNB regressions reported in Tables 2.10 and 2.11, specifications 0, 1a, 1b, and 1c.

Specification 0 tests the effects of individual, non-English home languages and educational environment on the total number of stickers shared. Specification 1 consolidates the individual home languages other than English into one indicator.

Non-English home language has a positive effect on the propensity not to share at all, compared to participants from exclusively English-speaking families. The increase in the propensity not to share any stickers over all trials is statistically significant for children from families with other or mixed home language. Using the consolidated indicator for all non-English home languages in specification 1 retains the positive effect on not sharing any stickers. However, it is not statistically significant (Table 2.10).

Considering the amount of stickers given, conditional on sharing at all, we can see in specification 0 of Table 2.11 that none of the home languages we evaluated other than English have a significantly positive effect on sharing. Except for the Mandarin-speaking children, all other participants shared less on average. The effect is strongest for children from Cantonese-speaking families, who shared on average 15% (more than two stickers) less in kindergarten (*grade*=0). The difference is significant at the 99% confidence level.

In specification 1 of the ZTNB regression, which does not distinguish the different non-English home languages, it can be seen that the effect of a non-English home language overall is significant at the 95% confidence level. A child in kindergarten with at least one other language than English spoken at home shared on average 13% ($\simeq$ two stickers) less than kindergarten children from exclusively English-speaking families.

The effect of age is statistically insignificant both for the propensity to share and the amount shared, conditional on a positive offer for children from English-speaking families. Specification 1a shows, however, that for participants with non-English home languages, sharing increases with age. The interaction effect between age and non-English home language is statistically significant at the 90% level for total stickers shared.

The variable *grade*, which is used to measure years of exposure to the Canadian public school environment, is insignificant for both the propensity to share nothing and for the amount shared, conditional on a positive offer. But, again, the interaction effect between with non-English home language is both significant (at the 95% confi-

dence level) and large for the amount of stickers shared (Table 2.11). For participants with at least one other home language, specifications 1b of the student-level ZTNB regression indicate that every year of schooling increases sticker offers in the dictator game by 11% ($\simeq$ 1.4 stickers over offers from this group in kindergarten).

Specification 1c of the hurdle model regressions include both age, grade, as well as their respective interaction effects with non-English home language. The effect of grade for children from non-English-speaking families on the amount shared, conditional on a positive offer remains significant at the 90% confidence level. Socialization in the educational environment for children from non-English-speaking families shows an even stronger influence when controlling for age specifically for this group. The effect of age for participants from non-English-speaking families, which indicates the duration of socialization in these families, changes sign and becomes negative. However, this effect is no longer statistically significant. The main effect for age and grade are both insignificant. The main effect of a non-English home language remains negative and significant at the 95% confidence level for the amount shared conditional on a positive offer (Table 2.11). For the propensity not to share, all language, age, and grade related effects in specification 1c are statistically insignificant (Table 2.10).

Specification 1c also controls for the effects of observed integration and perceived similarity. The number of classroom friends with Caucasian, East- or South-Asian phenotype, indicating objectively observable integration, has no statistically significant effect on the propensity to share, nor on the amount shared, conditional on a positive offer. Perceived similarity to the hypothetical others, however, is statistically significant at the 90% level for both probability to share and the amount offered. Every additional picture selected in the sorting task as 'like me' decreased the odds of sharing zero by 11% and increased the number of stickers shared by 1%. These percentages seem low compared to the other significant effects. However, perceived similarity is not a dichotomous variable. Between 0 and 12 pictures could be selected, with an observed mean of 3.8 (see Table 2.7).

The effects of home language, and grade for participants from non-English-speaking families retain their respective signs and remain significant in most trial and target level regressions reported in Tables 2.12-2.17. However, the influences of non-English home language and its interaction with grade become more and more relevant for the propensity to share, and less so for the amount offered. In most of the student-level regressions non-English home language has almost no significant effect on the odds of not sharing but a highly significant negative effect on the amounts offered. In the trial and target level regressions reported in Tables 2.16 and 2.17, however, non-English

home language has no significant influence on the amounts, but a significantly positive effect on the odds of not sharing at all. The reason is the different aggregation level of stickers shared. In the student level regressions, not sharing means that the participant kept all stickers for herself in all trials. The amount shared, conditional on at least one positive offer, adds up trials with positive offers and those with no sharing. The subsequent regressions de-compose this sum and indicate zero sharing on a trial (and picture) level.

Specifications 2 and 3 in the logit and ZTNB regressions reported in Tables 2.12 and 2.13, as well as the regressions reported in Tables 2.14-2.17 test for other influences on the sharing behaviour of the participants. In the student-level regressions, we find a significant influence on the amount of sharing for the order of testing (Tables 2.12 and 2.13, specification 3). Participants share significantly more often and share more stickers when the sorting task is done after the dictator game. The variable indicating that the dictator game was played first is significant at the 99% confidence level. A participant shared on average 10.0% ($\simeq$ 1.5 stickers) more when the dictator game came first. The odds of sharing nothing decrease by 61%.

Gender, the indicators for the sequential sharing treatment, and for the use of additional information on the hypothetical other are insignificant in the logit regression using student-level data, specification 2 and 3 (Table 2.12). In the ZTNB regression reported in Table 2.13, the use of additional information on the hypothetical others show a negative effect on the amount shared, which is significant at the 90% confidence level. Gender and sequential sharing show no significant influence. Relative height is insignificant for the amounts shared, conditional on a positive offer, but positive and large for the propensity to share nothing at all. for children shorter than average this effect is statistically significant at the 90% level.

Looking at the amount of stickers shared on a trial-by-trial basis (Tables 2.14 and 2.15), we find the number of trials to have a strong positive effect on the amount of stickers shared, conditional on sharing at all. Simultaneous sharing goes up by 6.9% ($\simeq$ 1 sticker), sequential sharing increases by 3.0% for each additional trial. The effect is statistically significant at the 99% confidence level in both simultaneous and sequential sharing treatment. Under the sequential sharing treatment, participants also shared significantly more often in later trials (significant at the 95% confidence level).

Trial-level data are also used to explore the effect of anonymity: Under the simultaneous sharing treatment, participants did not share less stickers, conditional on sharing at all. However, contrary to expectations, they shared more often. The

results reported in Table 2.14a show that increased anonymity decreases the odds of not sharing significantly (90% confidence level). The odds of keeping all stickers fall by 29%. Even more significant (at the 95% confidence level) is the anonymous treatment when looking at trial 3 in isolation in specification 2. Here the odds of not sharing even decrease by 43%. Under the sequential sharing treatment, we find no significant difference in the odds of sharing or the amounts of stickers shared between the open and the anonymous treatment.

The potential effects of identification with the hypothetical others are explored using trial and picture-level data. Providing information about the hypothetical others generally exerts no significant effect on the odds of sharing (Table 2.16), but has a negative influence on the amount of stickers shared, conditional on sharing at all (Table 2.17). The latter effect is significant at the 99% confidence level for specifications 1 and 3. It is insignificant in specification 4, which controls for perceived similarity.

The indicators for social distance based on exit interviews with the participants are all insignificant for the amount shared (specification 2 of the ZTNB regression). The odds of not sharing, however, increased by 23% when the participants reported that they know a child with the same name that was used to describe any of the hypothetical others. The other indicators of social distance were also insignificant in the logit regression reported in Table 2.16.

The fact that participant and hypothetical other are of the same phenotype is insignificant as well (specification 3, both logit and ZTNB model). However, when we include a measure of similarity to the hypothetical other as expressed by the participant, the outcome changes (specification 4). We find that the participant shared significantly more often and also shared more stickers the more pictures of the same phenotype as the hypothetical others are selected for the sorting task question "Who is like you?" Moreover, this effect is substantial. Up to four pictures with a given phenotype could be chosen as 'like me' in the sorting task. For each one that was selected, sharing went up by 4.7%; the odds of not sharing went down by 5.7%. The effect is highly statistically significant at the 99% confidence level.

The fact that the indicator for the sequential sharing treatment is highly significant in the target-level regressions reported in Tables 2.16 and 2.17 is a consequence of the experimental procedure. In the simultaneous sharing treatment, every participant was shown nine hypothetical others (three times three), while only six hypothetical others were shown in the sequential sharing treatment. However, the total endowment was

the same: 36 stickers. This led to more stickers being allocated to each hypothetical other, even if the total amount shared stayed the same.

All results reported in this subsection are robust when we control for tester effects. Only the level of significance is reduced for a few variables, due to the addition of 21 tester fixed effects.

2.6 Discussion

The regression results suggest that learning to share is part of the socialization of an individual in a specific cultural context. In particular, we can see the influence of exposure to the Canadian mainstream educational environment on the sharing behaviour of children from families with varying cultural backgrounds.

Kindergarten children from families with home languages other than English shared significantly less than their peers from exclusively English-speaking families. Exposure to the public school system compensates this difference. Sharing increases significantly for children from non-English-speaking families in grade 1 and 2.

Increased sharing also coincides with increased integration expressed by a higher number of classroom friends from one of the main phenotypes in the Greater Vancouver area, and higher perceived similarity to hypothetical others. However, the effect of socialization through longer exposure to the educational environment remains significant when controlling for objective and subjective levels of integration.

Based on these observations, we suggest that children are socialized to share at school. We assume that we can observe the socialization effect for children from non-English-speaking families because socialization at home and at school may differ.

We can not rule out that the increase in sharing by participants from non-English-speaking families in higher grades is partly due to red-shirting. However, the age distribution between the two types of participants show no significant difference. Participants from non-English-speaking families are not significantly older over all grades, which puts doubts on the extent of red-shirting.

For participants from exclusively English-speaking families, we find no significant effect of socialization in the educational environment. We do not find a significant increase in the total number of stickers shared with age for this group of children, either. Possibly, for children coming from exclusively English-speaking families, the socialization of this particular pro-social behaviour is already completed by the age of five. The values they acquired at home may coincide with what they learn at

school, so that exposure to the educational environment does not impact their sharing behaviour.

The correlations between home language, grade, and sharing stickers in the dictator game are significant and robust across different specifications of our econometric models, and under a comprehensive set of controls. The controls account for the main factors that are reported in the literature as influencing child behaviour in dictator games. We discuss them in detail below.

We find the influence of *age* on the amount of stickers shared to be insignificant. As for *gender*, we find no robust significant differences in the total number of stickers shared between boys and girls. As reported in Harbaugh et al. (2002), *Relative height* of a participant has a significant effect on offers on sharing, albeit in the opposite direction. Shorter participants share nothing significantly more often, but the number of stickers shared, based on a positive offer do not differ significantly. Taller than average participants also kept all sticker more often than participants with average height, even though this effect is just not statistically significant. A slightly different sample composition might have changed this significance. Therefore, we are cautious to reject Harbaugh et al. (2002)'s results.

We did find a significant and robust effect of the *order* in which we conducted our experiments. There are two reasons to expect that participants would share more if the dictator game is played first (see discussion in Smith 2007, p. 234ff). The first is based on the argument of 'other people's money' versus earned endowments. Participants view their endowments in the dictator game as earned and thus more rightfully theirs if the dictator game follows the long and relatively exhausting interview on attitudes. Therefore, participants might share less if the dictator game is played last.

The second argument for an order effect points to expected reciprocity. Participants could share more when they learn that the dictator game is only the first task and there is some interaction afterwards. They could anticipate some form of reciprocity to their behaviour in the dictator game and share more than when the dictator game was run as the last interaction with the tester. Unfortunately, these two reasons can not be distinguished with the experimental design used for our study.

The highly significant and positive effect of the number of repetitions of the experiment on the number of stickers shared can be attributed to a decreasing marginal utility of additional stickers. If there is a trade-off in utility between sharing and keeping stickers, the increased wealth of the participant in terms of stickers could have induced her or him to share more and more often in later trials of the dictator game.

We find no statistically significant negative effect of increased *anonymity* on offers in the dictator game that would be robust to variations in experimental procedures. To the contrary, children shared significantly more often under increased anonymity in the simultaneous sharing treatment.

However, we cannot rule out that the treatment was not effective and children were not convinced that their actions were indeed unobserved by the tester. In fact, we could not run double-blind treatments because we wanted to measure the behaviour of each participant and relate it to the participant's demographics and specific treatments. Therefore, only the joint hypothesis that anonymity increases sharing and our treatment effectively increases anonymity can be rejected. Further research is requested to verify how increased anonymity relates to the sharing behaviour of small children.

In general, *simultaneous* versus *sequential treatment* did not distort our results. Playing the dictator game three times with three hypothetical others simultaneously versus six times with one hypothetical other each did not have a significant effect on the total amount shared. The significant effect on the number of stickers shared with each hypothetical other is a logical consequence of the experimental design: A given participant played the dictator game with a total of nine or six hypothetical others, but with the same initial number of stickers between the simultaneous and sequential treatments.

Finally, we find evidence for identification with the hypothetical other to lead to increased sharing. However, the treatments we included in the dictator game to test for the two hypotheses of 'human face' and 'social distance' did not show the expected effect. To the contrary, simply providing additional information led to lower offers in the dictator game. Thus, the hypothesis that a 'human face' would increase sharing must be rejected based on our data.[12]

None of the indicators measuring decreased social distance had an effect that was significantly different from zero, either. However, we do not reject the hypothesis of an influence of 'social distance' in the same manner as we do for the 'human face' hypothesis. It is very likely that the characteristics introduced in the profiles were simply not relevant for the participants in the dictator games. But why does visible similarity in phenotype between dictator and hypothetical other have no effect? Regression results show that phenotype of the hypothetical other does matter. However,

[12] Of course it is possible that the additional infomation reduced the plausibility of the hypothetical other as a real human being because the information and the picture of the hypothetical other contradicted each other in the eyes of the participants.

there is a more subtle way a participant defines social distance to a hypothetical other that does not coincide with observable phenotype.

Only similarity, as perceived and declared by the participants themselves, was found to significantly increase the number of stickers offered in the dictator game. Thus, we find 'social distance' to be a significant factor in determining the sharing behaviour of children. However, trying to measure 'social distance' based on observable characteristics of the participant and the hypothetical other can be misleading. In our case, it seems that the properties introduced as additional information on the hypothetical others were ineffective in measuring social distance. In our context, self-declared similarity to the hypothetical others proved to be a significant predictor of sharing behaviour. Self-declared similarity did not fully coincide with equality in phenotype as classified by the teachers.

2.7 Conclusion

In this chapter, we examine whether altruistic behaviour of young children in dictator games differs significantly with their cultural background, indicated by the language(s) spoken in their respective families. We find that children from families in which English is not the native language spoken at home in Vancouver, British Columbia, Canada, share significantly less in kindergarten (age 5) than their peers with exclusively English-speaking parents.

We find strong evidence for altruism in dictator games to be shaped by socialization. In particular, exposure to the school environment seems to compensate the initial differences in sharing behaviour. Children with home language(s) other than English shared significantly more sticker in grades 1 and 2 than in kindergarten, while the effect was absent for their peers from exclusively English-speaking families.

Our data set allows us to control for a multitude of other possible determinants of the behaviour of children in dictator games. Most importantly for the discussion of socialization, we can include indicators for the level of integration. While the number of classroom friends of Caucasian, East and South Asian phenotype has no statistically significant effect on sharing with hypothetical others of these phenotypes, the participants' self-declared similarity to children of these main phenotypes has a strong positive effects on the level of sharing. The language spoken at home and exposure to socialization in a public school setting remain the strongest explanatory variables for the differences in offers in the dictator games. We conclude that a major driver of pro-social behaviour of children is their socialization in a specific cultural

context. Moreover, we are able to illuminate two specific channels of socialization. A participant's family background and the exposure to a public school environment are the strongest determinants of the level of altruism for children aged 5 to 8.

In addition we see the influence of 'social distance'. However, 'social distance' in our case is only a significant driver of behaviour when understood as the self-declared similarity to the hypothetical other rather than some externally constructed measure based on phenotype or similarities in preferences. Our participants shared more if they felt similar to the hypothetical others, not necessarily if they looked or behaved the same.

Using our rich sample derived from experiments with 1,018 primary school children, we investigated other possible influences on the sharing behaviour of young children, which have been reported as significant in the literature. We find clear evidence of an *order effect*, which is consistent with the notion that less of an endowment is given away when it was regarded as earned by the dictator rather than just 'other people's money' appearing for no reason. Our participants also shared significantly more with every trial that the experiment was repeated. This suggests that children react as expected to the effect of wealth when evaluating trade-offs in economic behaviour. The reaction of our participants to increased anonymity is surprising and requires further study, including refinements to the experimental procedures.

We find no conclusive and robust evidence for differences in altruism to be explained by biological characteristics. *Age*, *gender*, and *relative height* did not influence the amounts shared significantly. We did find *relative height* to be significant for the propensity to share, even when we control for the main determinants of socialization. However, the significance level is low and *age* and *gender* are not significant in this respect, either. Thus, in our study, socialization prevails. Of course, our study only covers a small fraction of a child's ontogenesis between the ages of 5 and 8. Universal (genetic) developmental influences at later and, in particular, earlier ages, can not be ruled out here. Further empirical research is needed to disentangle these two effects – the dominance of nature or nurture on the development of altruism.

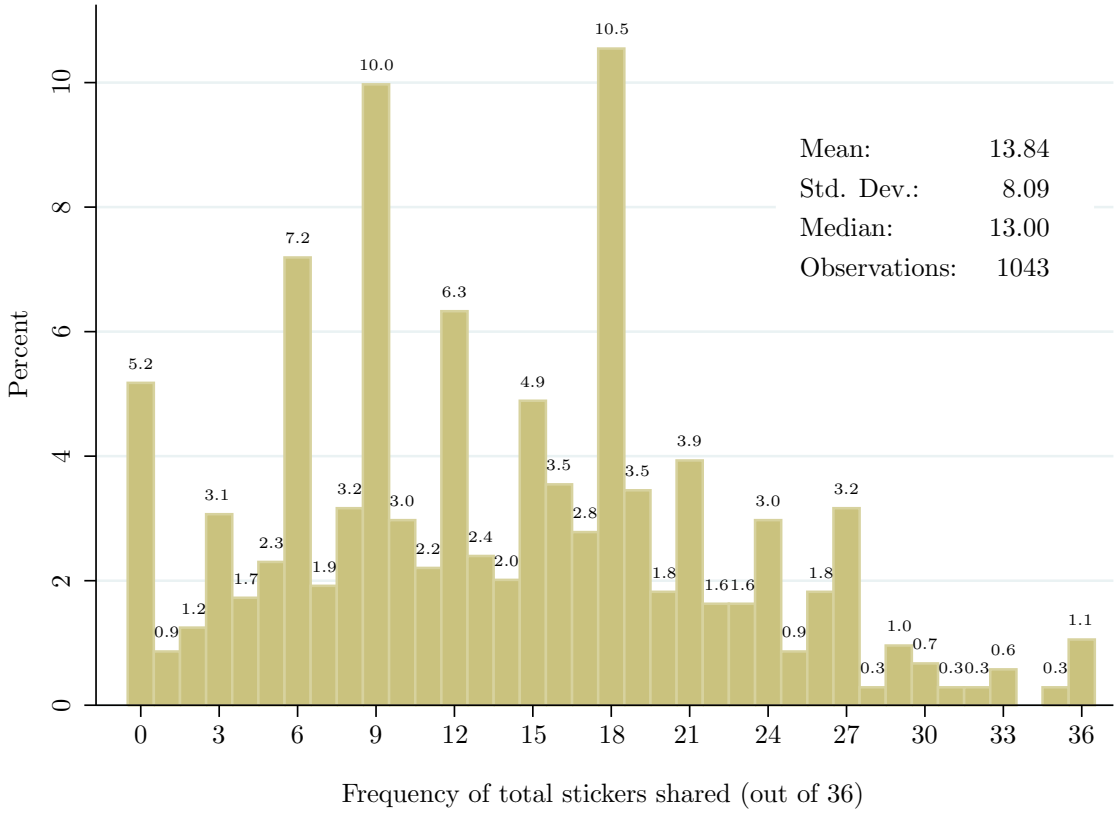
Figures

Figure 2.1: Pictures of children as 'hypothetical others'

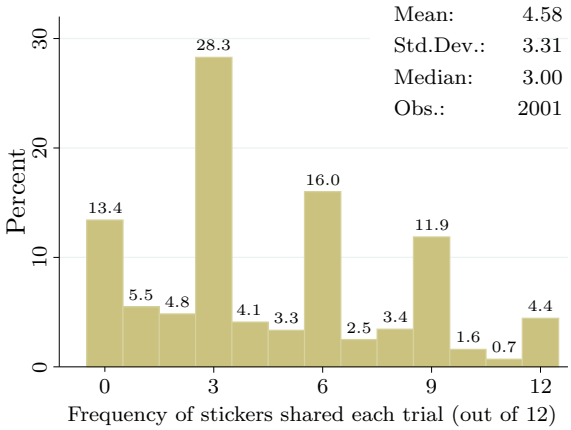
Figure 2.2: Plain, multicolored stickers as endowments in dictator games

Figure 2.3: Summary statistics and histograms of stickers shared in the dictator game

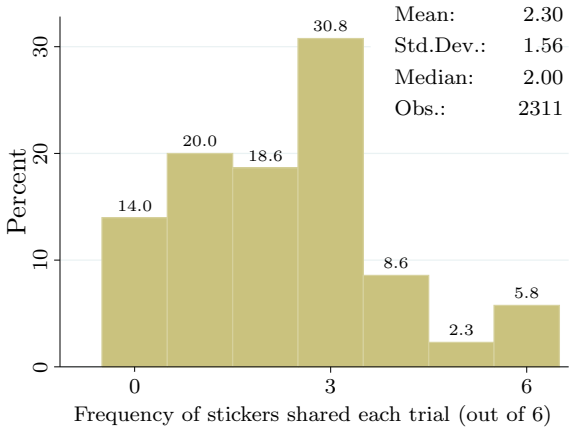
(a) Total number of stickers shared over all trials and treatments



(b) Sharing per trial (simultaneous treatment)



(c) Sharing per trial (sequential treatment)



Tables

Table 2.1: Sample characteristics

	Frequency	Percent		Frequency	Percent
Age			**Gender**		
5	308	29.5	Male	544	52.2
6	399	38.3	Female	499	47.8
7	263	25.2			
8[a]	71	6.8	**Home Language**[b]		
(missing)	2	0.2	English	484	46.4
			Cantonese	98	10.2
Phenotype[b]			Chinese[c]	88	9.2
Caucasian	328	31.5	Mandarin	35	3.7
East Asian	306	29.3	Korean	9	0.9
- Chinese	*290*	*27.8*	Japanese	6	0.6
- Japanese	*6*	*0.6*	Punjabi	34	3.3
- Korean	*10*	*1.0*	Hindi	4	0.4
South Asian[d]	74	7.1	Gujarati	4	0.4
SE Asian	99	9.5	Tamil	2	0.2
- Vietnamese	*36*	*3.5*	Bengali	1	0.1
- Phillipino	*63*	*6.0*	Vietnamese	32	3.1
Other[e]	159	15.2	Tagalog[f]	25	2.6
(missing)	77	7.4	Spanish	13	1.3
			Russian	4	0.4
Grade			Somali	2	0.2
Kindergarten	431	41.3	French	1	0.1
Grade 1	350	33.6	Other/mixed	116	12.1
Grade 2	219	21.0	(missing)	85	8.2
(missing)[g]	43	4.1			
			Total	1043	100.0

[a] Includes 3 children aged 9.

[b] Teacher information.

[c] Not specified further.

[d] Indian subcontinent, including Pakistan, Bangladesh, and Sri Lanka.

[e] Black, Hispanic/Latin American, Aboriginal/First Nation, Middle Eastern and mixed phenotype.

[f] Main language spoken in the Philippines and basis for its official language, Filipino.

[g] For some mixed-grade classrooms, the individual grade of each participant was not recorded. In total, 70 children (7%) went into mixed K/G1, and 59 (6%) into mixed G1/G2 classes, while 409 children (39%) went into pure kindergarten groups, 301 (29%) into Grade 1, and 204 (20%) into Grade 2 classes.

Table 2.2: English home language and phenotype

	Phenotype (grouped, teacher information)						
Home Language[a]	Caucasian	East Asian	South Asian	SE Asian	Other/ Mixed	(missing)	Total
English	302	46	16	15	105	0	484
	(92.1)	(15.0)	(21.6)	(15.2)	(66.0)	(0.0)	(46.4)
Other	26	254	57	84	53	0	474
	(7.9)	(83.0)	(77.0)	(84.8)	(33.3)	(0.0)	(45.4)
(missing)	0	6	1	0	1	77	85
	(0.0)	(2.0)	(1.4)	(0.0)	(0.7)	(100.0)	(8.2)
Total	328	306	74	99	159	77	1043
	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

[a] Teacher information (frequency and percentage in brackets).

Table 2.3: Reported age by grade and phenotype, as well as by home language (grouped)

(a) Kindergarten

	Phenotype (grouped, teacher information)						**Home Language**			
Age[a]	Caucasian	East Asian	South Asian	S. East Asian	Other/ mixed	(msg)	English[b]	Other[c]	(msg)	Total
5	98	89	17	40	43	15	155	131	16	302
	(74.8)	(59.3)	(65.4)	(81.6)	(82.7)	(65.2)	(76.0)	(65.5)	(59.3)	(70.1)
6	32	53	8	8	8	7	45	61	10	116
	(24.4)	(35.3)	(30.8)	(16.3)	(15.4)	(30.4)	(22.0)	(30.5)	(37.0)	(26.9)
7	1	7	1	1	1	1	4	7	1	12
	(0.8)	(4.7)	(3.8)	(2.1)	(1.9)	(4.4)	(2.0)	(3.5)	(3.7)	(2.8)
(msg)	0	1	0	0	0	0	0	1	0	1
	(0.0)	(0.7)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.5)	(0.0)	(0.2)
Total	131	150	26	49	52	23	204	200	27	431
	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

(b) Grade 1

	Phenotype (grouped, teacher information)						**Home Language**			
Age[a]	Caucasian	East Asian	South Asian	S. East Asian	Other/ mixed	(msg)	English[b]	Other[c]	(msg)	Total
5	1	1	0	0	1	0	1	2	0	3
	(0.8)	(1.2)	(0.0)	(0.0)	(1.4)	(0.0)	(0.5)	(1.4)	(0.0)	(0.9)
6	90	65	16	15	54	17	130	106	21	257
	(70.3)	(79.3)	(64.0)	(62.5)	(74.0)	(94.4)	(70.3)	(74.1)	(95.5)	(73.4)
7	35	16	9	9	18	1	52	35	1	88
	(27.3)	(19.5)	(36.0)	(37.5)	(24.6)	(5.6)	(28.1)	(24.5)	(4.5)	(25.1)
8	2	0	0	0	0	0	2	0	0	2
	(1.6)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(1.1)	(0.0)	(0.0)	(0.6)
Total	128	82	25	24	73	18	185	143	22	350
	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

(c) Grade 2

	Phenotype (grouped, teacher information)						**Home Language**			
Age[a]	Caucasian	East Asian	South Asian	S. East Asian	Other/ mixed	(msg)	English[b]	Other[c]	(msg)	Total
5	0	1	0	0	0	0	0	1	0	1
	(0.0)	(2.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(1.0)	(0.0)	(0.5)
6	3	2	1	1	5	0	6	6	0	12
	(5.0)	(3.9)	(5.3)	(4.6)	(16.1)	(0.0)	(7.5)	(5.8)	(0.0)	(5.5)
7	41	34	11	17	16	23	53	66	23	142
	(68.3)	(66.6)	(57.9)	(77.2)	(51.6)	(63.9)	(66.2)	(64.1)	(63.9)	(64.7)
8	16	14	7	4	10	12	21	30	12	63
	(26.7)	(27.5)	(36.8)	(18.2)	(32.3)	(33.3)	(26.3)	(29.1)	(33.3)	(28.8)
(msg)	0	0	0	0	0	1	0	0	1	1
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(2.8)	(0.0)	(0.0)	(2.8)	(0.5)
Total	60	51	19	22	31	36	80	103	36	219
	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

[a] Age as reported by participant (frequency, percentage in brackets). Grouped: 5 = 5.0 to 5.9, 6 = 6.0 to 6.9 etc.

[b] Participants from exclusively English-speaking families.

[c] Participants from families speaking at least one language other than English at home.

Table 2.4: Test for difference in age distribution between English and Non-English home language participants

Two-sample Kolmogorov-Smirnov test for equality of distributions

Smaller group	**D**	**P-Value**	**Exact**
No (Non-English home language)	0.0000	1.000	
Yes (English home language)	-0.0663	0.122	
Combined K-S	0.0663	0.244	0.229

Table 2.5: Test for difference in average age between English and Non-English home language participants by grade

(a) Equality of means test - Kindergarten

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
No (Non-English home language)	199	5.404523	.038757	.5467347	5.328093	5.480952
Yes (English home language)	204	5.320098	.0336987	.4813139	5.253654	5.386542
combined	403	5.361787	.0256914	.5157507	5.31128	5.412293
diff = mean(No) - mean(Yes)		.0844246	.0513586		-.0165482	.1853974
$H_a : diff < 0$		$H_a : diff \neq 0$			$H_a : diff > 0$	
$Pr(T < t) = 0.9495$		$Pr(\|T\| > \|t\|) = 0.1010$			$Pr(T > t) = 0.0505$	

(b) Equality of means test - Grade 1

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
No (Non-English home language)	143	6.28951	.0381245	.4559023	6.214146	6.364875
Yes (English home language)	185	6.323243	.0379831	.516626	6.248305	6.398182
combined	328	6.308537	.0270907	.4906343	6.255242	6.361831
diff		-.0337328	.0538163		-.1396108	.0721453
$H_a : diff < 0$		$H_a : diff \neq 0$			$H_a : diff > 0$	
$Pr(T < t) = 0.2656$		$Pr(\|T\| > \|t\|) = 0.5312$			$Pr(T > t) = 0.7344$	

(c) Equality of means test - Grade 2

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
No (Non-English home language)	103	7.271845	.0600529	.6094706	7.15273	7.390959
Yes (English home language)	80	7.25125	.0615978	.5509471	7.128643	7.373857
combined	183	7.262842	.0431062	.5831297	7.177789	7.347894
diff		.0205947	.086027		-.149177	.1903664
$H_a : diff < 0$		$H_a : diff \neq 0$			$H_a : diff > 0$	
$Pr(T < t) = 0.5945$		$Pr(\|T\| > \|t\|) = 0.8111$			$Pr(T > t) = 0.4055$	

Table 2.6: English home language and number of friends in the class

	Caucasian, East or South Asian Classroom Friends (total number, teacher information)[a]							
Home Language[b]	0	1	2	3	4	5	(missing)	Total
English	20	85	136	123	80	34	6	484
	(4.1)	(17.6)	(28.1)	(25.4)	(15.5)	(7.0)	(1.2)	(100.0)
Other	34	97	129	102	52	37	23	474
	(7.2)	(20.5)	(27.2)	(21.5)	(11.0)	(7.8)	(4.9)	(100.0)
(missing)	1	3	0	3	4	87	66	85
	(1.2)	(3.5)	(0.0)	(3.5)	(4.7)	(9.4)	(77.7)	(100.0)
Total	55	185	265	228	136	79	95	1043
	(5.3)	(17.4)	(25.4)	(21.9)	(13.0)	(7.6)	(9.1)	(100.0)

[a] Frequency and percentage in brackets.
[b] Teacher information.

Table 2.7: Perceived similarity to hypothetical others by phenotype

	Phenotype (grouped, teacher information)						
Like You Total[a]	Caucasian	East Asian	South Asian	SE Asian	Other/ mixed	(missing)	Total
0	32	40	6	8	17	8	111
	(9.8)	(13.1)	(8.1)	(8.1)	(10.7)	(10.4)	(10.6)
1	38	28	7	10	21	5	109
	(11.6)	(9.2)	(9.5)	(10.1)	(13.2)	(6.5)	(10.5)
2	48	39	7	7	18	19	138
	(14.6)	(12.8)	(9.5)	(7.1)	(11.3)	(24.7)	(13.2)
3	45	36	5	8	15	7	116
	(13.7)	(11.8)	(6.8)	(8.1)	(9.4)	(9.1)	(11.1)
4	42	48	11	9	19	4	133
	(12.8)	(15.7)	(14.9)	(9.1)	(12.0)	(5.2)	(12.8)
5	24	20	6	11	13	8	82
	(7.3)	(6.5)	(8.1)	(11.1)	(8.2)	(10.4)	(7.9)
6	24	18	8	5	12	5	72
	(7.3)	(5.9)	(10.8)	(5.1)	(7.6)	(6.5)	(6.9)
7	6	9	3	3	4	3	28
	(1.8)	(2.9)	(4.1)	(3.0)	(2.5)	(3.9)	(2.7)
8	9	6	2	3	3	0	23
	(2.7)	(2.0)	(2.7)	(3.0)	(1.9)	(0.0)	(2.2)
9	4	1	2	1	4	1	13
	(1.2)	(0.3)	(2.7)	(1.0)	(2.5)	(1.3)	(1.3)
10	1	2	2	3	0	0	8
	(0.3)	(0.7)	(2.7)	(3.0)	(0.0)	(0.0)	(0.8)
11	1	2	0	1	4	3	11
	(0.3)	(0.7)	(0.0)	(1.0)	(2.5)	(3.9)	(1.1)
12	14	14	3	9	4	3	47
	(4.3)	(4.6)	(4.1)	(9.1)	(2.5)	(3.9)	(4.5)
(missing)	40	43	12	21	25	11	152
	(12.2)	(14.1)	(16.2)	(21.2)	(15.7)	(14.3)	(14.6)
Total	328	306	74	99	159	77	1043
	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)	(100.0)

[a] Times phenotype of hypothetical other(s) selected in sorting task question: "Who is like you?"Frequency, and percentage in brackets.

Table 2.8: Additional information, randomly attributed to hypothetical others

Profile	**Name**		**Fav. ice cream**	**Fav. activity/sports**	
	Girls	*Boys*		*Girls*	*Boys*
A	Sarah	Jack	Chocolate	Jump rope	Baseball
B	Jessica	Michael	Vanilla	Dodge ball	Soccer
C	Karen	Chris	Strawberry	Playground	Hockey

Table 2.9: Data sets used in regressions

		Number of	
Data set	**Dependent variable**	**Observations**	**Unique participants**
Student level	Total stickers shared by participant in all trials	1,043[a]	1,018[b]
Trial level (simultaneous)[c]	Total stickers shared per trial and participant	2,008	659[d]
Trial level (sequential) [e]	Total stickers shared per trial and participant	2,311	387[f]
Trial x picture	Stickers shared with a given hypothetical other per trial and participant	8,835	1,036[g]

[a] Completed all trials of the dictator game.

[b] Children who participated in both 2007 and 2008 counted only once.

[c] All children tested in simultaneous sharing treatment.

[d] Includes 16 children under simultaneous sharing treatment in both 2007 and 08, one of them for only 1 out of 3 trials.

[e] All children tested in sequential sharing treatment.

[f] Includes 10 participants in 2008 who completed some or all trials under simultaneous sharing treatment in 2007.

[g] Includes 18 participants who completed some, but not all trials.

Table 2.10: Logit regression of sharing zero over all trials (main results)

Logit, Dependent variable: Indicator for zero stickers shared over all trials Standard errors clustered by class id				
	Specification number			
Independent variables	0	1a	1b	1c
Socialization[a]				
Chinese[b]	0.388			
	(0.517)			
Mandarin	0.394			
	(0.765)			
Cantonese	0.041			
	(0.562)			
Tagalog[c]	_[d]			
Punjabi	0.862			
	(0.648)			
Vietnamese	_[d]			
Other/mixed	0.822**			
	(0.327)			
Home language not English		0.570	0.652	0.288
		(0.436)	(0.406)	(0.514)
Age[e]		−0.402		−0.835
		(0.295)		(0.619)
Age[e] x Non-English		−0.272		0.321
		(0.396)		(0.781)
Grade[f]			−0.253	0.529
			(0.345)	(0.683)
Grade[f] x Non-English			−0.364	−0.468
			(0.456)	(0.858)
Perceived similarity to hypothetical other[g]				−0.118*
				(0.064)
Caucasian, East or South Asian classroom friends[h]				−0.167
				(0.137)
Constant	−3.198***	−2.825***	−3.050***	−1.952***
	(0.234)	(0.333)	(0.321)	(0.543)
N	986	957	915	769
ll	−205.70	−173.22	−172.75	−137.64
Significance levels: * 90%, ** 95%, *** 99%				

[a] Home language, exposure to school environment, and integration measures.

[b] Not specified further.

[c] Main language spoken in the Philippines and basis for its official language, Filipino.

[d] Variable dropped because it predicts outcome perfectly (participant always shared non-zero amount).

[e] Normalised: Actual age - minimum age of 5.

[f] Grade=0 for Kindergarten children.

[g] Times phenotype of hypothetical other(s) selected in sorting task question: "Who is like you?"

[h] Teacher information, total of 0 to 5 friends possible.

Table 2.11: Zero truncated negative binomial regression of stickers shared conditional on a positive offer (main results)

ZTNB, Dependent variable: Total number of stickers shared over all trials, conditional on a positive offer Standard errors clustered by class id				
	Specification number			
Independent variables	0	1a	1b	1c
Socialization[a]				
Chinese[b]	−0.060			
	(0.065)			
Mandarin	0.058			
	(0.105)			
Cantonese	−0.166***			
	(0.063)			
Tagalog[c]	−0.089			
	(0.127)			
Punjabi	−0.111			
	(0.098)			
Vietnamese	−0.062			
	(0.078)			
Other/mixed	−0.006			
	(0.049)			
Home language not English		−0.135**	−0.147***	−0.128**
		(0.053)	(0.050)	(0.058)
Age[d]		−0.019		−0.006
		(0.025)		(0.059)
Age[d] x Non-English		0.060*		−0.051
		(0.034)		(0.074)
Grade[e]			−0.022	0.006
			(0.033)	(0.070)
Grade[e] x Non-English			0.103**	0.139*
			(0.042)	(0.084)
Perceived similarity to hypothetical other[f]				0.013*
				(0.007)
Caucasian, East or South Asian classroom friends[g]				0.008
				(0.015)
Constant	2.705***	2.726***	2.718***	2.638***
	(0.025)	(0.038)	(0.036)	(0.069)
$ln(\alpha)$ Constant	−1.462***	−1.466***	−1.436***	−1.416***
	(0.058)	(0.061)	(0.060)	(0.067)
N	989	913	871	734
ll	−3371.43	−3107.34	−2969.06	−2505.03
Significance levels: * 90%, ** 95%, *** 99%				

[a] Home language, exposure to school environment, and integration measures.

[b] Not specified further.

[c] Main language spoken in the Philippines and basis for its official language, Filipino.

[d] Normalised: Actual age - minimum age of 5.

[e] Grade=0 for Kindergarten children.

[f] Times phenotype of hypothetical other(s) selected in sorting task question: "Who is like you?"

[g] Teacher information, total of 0 to 5 friends possible.

Table 2.12: Logit regression of sharing zero over all trials (other variables)

Logit, Dependent variable: Indicator for zero stickers shared over all trials

Standard errors clustered by class id

	Specification number		
Independent variables	1b	2	3
Socialization[a]			
Home language not English	0.652	0.617	0.705*
	(0.406)	(0.482)	(0.409)
Grade[f]	−0.253	−0.297	−0.239
	(0.345)	(0.360)	(0.346)
Grade[f] x Non-English	−0.364	−0.493	−0.405
	(0.456)	(0.499)	(0.460)
Demographics			
Female		0.025	
		(0.352)	
Relative height[b]			
Shorter than avg		0.812*	
		(0.418)	
Taller than avg		0.730	
		(0.445)	
Treatments			
Dictator game first			−0.932***
			(0.329)
Sequential sharing			0.045
			(0.331)
Additional information			0.037
on hypothetical others			(0.313)
Constant	−3.050***	−3.335***	−2.690***
	(0.321)	(0.434)	(0.388)
N	915	759	915
ll	−172.75	−135.64	−168.42

Significance levels: * 90%, ** 95%, *** 99%

[a] Home language, exposure to school environment, and integration measures.

[b] Relative to classmates, binary indicators, baseline is average height (dummy left out).

Table 2.13: Zero truncated negative binomial regression of stickers shared conditional on a positive offer (other variables)

ZTNB, Dependent variable: Total number of stickers shared over all trials, conditional on a positive offer
Standard errors clustered by class id

	Specification number		
Independent variables	1b	2	3
Socialization[a]			
Home language not English	−0.147***	−0.141***	−0.151***
	(0.050)	(0.051)	(0.050)
Grade[e]	−0.022	−0.031	−0.025
	(0.033)	(0.033)	(0.035)
Grade[e] x Non-English	0.103**	0.105***	0.105**
	(0.042)	(0.040)	(0.044)
Demographics			
Female		0.047	
		(0.038)	
Relative height[b]			
Shorter than avg		0.027	
		(0.054)	
Taller than avg		0.055	
		(0.054)	
Treatments			
Dictator game first			0.096***
			(0.037)
Sequential sharing			−0.004
			(0.043)
Additional information			−0.057*
on hypothetical others			(0.031)
Constant	2.718***	2.698***	2.695***
	(0.036)	(0.047)	(0.046)
$ln(\alpha)$			
Constant	−1.436***	−1.513***	−1.450***
	(0.060)	(0.060)	(0.059)
N	871	724	871
ll	−2969.06	−2467.73	−2964.53

Significance levels: * 90%, ** 95%, *** 99%

[a] Home language, exposure to school environment, and integration measures.
[b] Relative to classmates, binary indicators, baseline is average height (dummy left out).

Table 2.14: Trial level regressions - hurdle model for simultaneous sharing

(a) Logit: Zero stickers shared

Logit, Dep. var.: Zeros shared per trial
Standard errors clustered by part. id

	Specification number	
Indep. var.	1	2
Socialization[a]		
Non-Eng. home	0.545***	0.551***
	(0.189)	(0.189)
Grade[b]	−0.208	−0.208
	(0.159)	(0.159)
Grade[b] x N-Eng	−0.311	−0.310
	(0.217)	(0.217)
Treatments		
Dictator first	−0.945***	−0.943***
	(0.152)	(0.152)
Profiles used	0.164	0.166
	(0.145)	(0.145)
Trial	−0.158	−0.116
	(0.098)	(0.104)
Anonymity		
Anon. sharing	−0.336*	
	(0.183)	
Trial 2 anon.		−0.211
		(0.207)
Trial 3 anon.		−0.564**
		(0.272)
Constant	−1.229***	−1.305***
	(0.241)	(0.251)
N	1768	1768
ll	−643.77	−643.10

Significance levels: * 90%, ** 95%, *** 99%

(b) ZTNB: Total sharing (pos. offers)

ZTNB, Dep. var.: Stickers shared per trial, conditional on a positive offer
Standard errors clustered by part. id

	Specification number	
Indep. var.	1	2
Socialization[a]		
Non-Eng. home	−0.133**	−0.132**
	(0.067)	(0.067)
Grade[b]	0.028	0.028
	(0.041)	(0.041)
Grade[b] x N-Eng	0.113*	0.113*
	(0.060)	(0.060)
Treatments		
Dictator first	0.062	0.062
	(0.046)	(0.046)
Profiles used	−0.079*	−0.079*
	(0.046)	(0.046)
Trial	0.061***	0.067***
	(0.014)	(0.016)
Anonymity		
Anon. sharing	0.004	
	(0.022)	
Trial 2 anon.		0.018
		(0.033)
Trial 3 anon.		−0.017
		(0.043)
Constant	1.511***	1.500***
	(0.062)	(0.063)
$ln(\alpha)$ const.	−1.792***	−1.793***
	(0.104)	(0.104)
N	1540	1540
ll	−3705.94	−3705.94

Significance levels: * 90%, ** 95%, *** 99%

[a] Home language, exposure to school environment, and integration measures.

[b] Grade=0 for Kindergarten children.

Table 2.15: Trial level regressions - hurdle model for sequential sharing

(a) Logit: Zero stickers shared

Logit, Dep. var.: Zeros/trial Errors clustered by pid	
	Spec. no.
Indep. var.	1
Socialization[a]	
Non-Eng. home	0.953***
	(0.199)
Grade[b]	0.210
	(0.128)
Grade[b]x N-Eng	−0.482***
	(0.163)
Treatments	
Dictator game first	−0.028
	(0.134)
Profiles used	0.121
	(0.132)
Trial	−0.076**
	(0.038)
Anonymity	
Anon. sharing	0.157
	(0.129)
Constant	−2.138***
	(0.237)
N	2012
ll	−808.00
Sign.: * 90%, ** 95%, *** 99%	

(b) ZTNB: Total sharing (pos. offers)

ZTNB, Dep. var.: Stickers/trial, conditional on pos. offers Errors clustered by part.id	
	Spec. no.
Indep. var.	1
Socialization[a]	
Non-Eng. home	−0.060
	(0.093)
Grade[b]	−0.099
	(0.052)
Grade[b]x N-Eng	0.055
	(0.078)
Treatments	
Dictator game first	0.135**
	(0.064)
Profiles used	−0.007
	(0.063)
Trial	0.023***
	(0.008)
Anonymity	
Anon. sharing	−0.028
	(0.024)
Constant	0.840***
	(0.084)
$ln(\alpha)$ const.	−21.075
	(.)
N	1725
ll	−2837.54
Sign.: * 90%, ** 95%, *** 99%	

[a] Home language, exposure to school environment, and integration measures.

[b] Grade=0 for Kindergarten children.

Table 2.16: Logit regressions of zero stickers shared per trial and picture

Logit, Dependent variable: Zero stickers shared per trial & picture Standard errors clustered by participant id				
	Specification number			
Independent variables	1	2	3	4
Socialization[a]				
Home language not English	0.561***	0.543***	0.561***	0.407***
	(0.080)	(0.150)	(0.080)	(0.087)
Grade[b]	−0.086	−0.296***	−0.086	−0.169**
	(0.061)	(0.105)	(0.061)	(0.067)
Grade[b] x Non-English	−0.242***	0.010	−0.242***	−0.076
	(0.080)	(0.141)	(0.080)	(0.090)
Treatments				
Dictator game first	−0.484***	−0.707***	−0.484***	−0.472***
	(0.059)	(0.104)	(0.059)	(0.065)
Sequential sharing	−0.229***	−0.271**	−0.229***	−0.344***
	(0.078)	(0.132)	(0.078)	(0.091)
Trial	−0.114***	−0.139***	−0.114***	−0.135***
	(0.027)	(0.046)	(0.027)	(0.031)
Profiles used	0.061		0.061	0.088
	(0.059)		(0.059)	(0.064)
Identification				
Knows name		0.227**		
		(0.104)		
Likes same ice cream		−0.024		
		(0.113)		
Likes same sports/activity		0.139		
		(0.134)		
Same phenotype as hypoth. other			−0.002	−0.021
			(0.069)	(0.076)
Perceived similarity to phenotype[c]				−0.064**
				(0.026)
Constant	−1.012***	−0.852***	−1.012***	−0.827***
	(0.091)	(0.153)	(0.092)	(0.104)
N	7684	2633	7684	6402
ll	−3651.24	−1220.44	−3651.24	−3077.15
Significance levels: * 90%, ** 95%, *** 99%				

[a] Home language, exposure to school environment, and integration measures.

[b] Grade=0 for Kindergarten children.

[c] Times phenotype of hypothetical other(s) selected in sorting task question: "Who is like you?"

Table 2.17: ZTNB regressions of total stickers shared per trial and picture, conditional on a positive offer

ZTNB, Dependent variable: Stickers shared per trial and picture, conditional on a positive offer Standard errors clustered by participant id				
	Specification number			
Independent variables	1	2	3	4
Socialization[a]				
Home language not English	−0.079	−0.072	−0.079	−0.131*
	(0.085)	(0.123)	(0.085)	(0.079)
Grade[b]	−0.022	−0.023	−0.022	−0.014
	(0.046)	(0.065)	(0.046)	(0.047)
Grade[b] x Non-English	0.113*	0.040	0.113*	0.124*
	(0.067)	(0.099)	(0.067)	(0.066)
Treatments				
Dictator game first	0.145***	0.101	0.145***	0.086*
	(0.052)	(0.076)	(0.052)	(0.051)
Sequential sharing	0.548***	0.439***	0.549***	0.395***
	(0.056)	(0.079)	(0.056)	(0.054)
Trial	0.036***	0.039***	0.036***	0.044***
	(0.009)	(0.012)	(0.009)	(0.009)
Profiles used	−0.145***		−0.145***	−0.055
	(0.052)		(0.053)	(0.051)
Identification				
Knows name		0.004		
		(0.050)		
Likes same ice cream		−0.019		
		(0.042)		
Likes same sports/activity		−0.057		
		(0.056)		
Same phenotype as hypoth. other			0.009	−0.046
			(0.029)	(0.029)
Perceived similarity to phenotype[c]				0.050***
				(0.016)
Constant	0.252***	0.316***	0.250***	0.273***
	(0.078)	(0.096)	(0.079)	(0.078)
$ln(\alpha)$ constant	−1.863***	−20.779	−1.863***	−4.317*
	(0.255)	(.)	(0.255)	(2.203)
N	6204	2135	6204	5149
ll	−9224.09	−3025.35	−9224.04	−7210.97
Significance levels: * 90%, ** 95%, *** 99%				

[a] Home language, exposure to school environment, and integration measures.

[b] Grade=0 for Kindergarten children.

[c] Times phenotype of hypothetical other(s) selected in sorting task question: "Who is like you?"

Chapter 3

Ethnic Identity and Discrimination among Children

3.1 Introduction

A large empirical literature in social psychology and economics demonstrates that individuals tend to favour members of groups with whom they associate themselves. Social identity theory, which posits that individuals place themselves and others in groups and make comparisons across groups, provides a conceptual framework for understanding intergroup relations and, particularly, this tendency for individuals to favour members of "in-groups" with whom they identify (Tajfel & Turner 1986).

Much of the evidence for what has been referred to as the in-group bias effect has come from experiments in which behaviour is compared across neutral conditions and conditions where identity is made more salient, either by inducing artificial identities in the "minimal group paradigm" (Tajfel & Turner 1986), or by priming natural identities. Recent examples of minimal group studies involving economic games include Chen & Li (2009), who find greater charity towards and less envy of in-group members, a stronger tendency to forgive and weaker tendency to punish bad intentions of in-group members, and a greater likelihood of choosing social-welfare-maximizing actions when participants are matched with an in-group member. Studies that involve priming natural social identities include Shih, Pittinsky & Ambady (1999), Benjamin, Choi & Fisher (2010), Benjamin, Choi & Strickland (2010), as well as many studies in social psychology. Evidence that in-group bias is pervasive with respect to a wide variety of naturally-occurring groupings comes from studies that find preferential treatment towards in-groups in the context of real, unprimed social groupings,

including college fraternities (Kollock 1998), tribes in Papua New Guinea (Bernhard, Fischbacher & Fehr 2006), Swiss Army platoons (Goette, Huffman & Meier 2006), schools (Fehr et al. 2008), and groups defined by a variety of personal characteristics (Ben-Ner, McCall, Stephane & Wang 2009). Racial and ethnic bias has also been found in dictator games among South African subjects (Burns 2010) and among Israeli subjects (Fershtman & Gneezy 2001).

This study explores in-group bias in children's altruistic behaviour. Studies involving children can inform our understanding of the developmental processes associated with in-group bias, and may shed light on the evolutionary forces that have shaped these social preferences (Fehr et al. 2008). Previous research provides mixed evidence about whether children's altruistic behaviour is characterized by in-group bias. Fehr et al. (2008) find that children aged three to eight are more likely to choose an egalitarian allocation when dividing resources between themselves and an in-group member (defined as a child who attends the same pre-school, daycare or school) than with an out-group member. McGillicuddy-de Lisi, Watkins & Vinchur (1994) find no evidence of in-group bias in allocations by young elementary school children to black and white story characters. Goeree et al. (2010) find that dictator offers to schoolmates among girls aged ten to twelve are explained by social distance to the recipient, as measured by school friendship networks; however, observable characteristics of the recipient, including race, play little or no direct role in determining their allocations. Two studies that use the minimal group paradigm with children also produce mixed results. Spielman (2000) finds that Kindergarten children allocate more resources to in-group than out-group members only when they are competitively primed; Gummerum, Takezawa & Keller (2009) find that sixth grade students exhibit in-group bias in a series of economic games, but second grade children do not.

We examine children's altruistic behaviour towards children from phenotypically different ethnic groups. Over 430 children between the ages of five and eight, belonging to three different ethnic categories (White, East Asian and South Asian) and attending public school in Vancouver, Canada participated in our study in 2007 and 2008. We engaged these children in a series of activities that draw from both social psychology and experimental economics and are designed to reveal their evaluations of other children from each of these three ethnic categories (i.e. White, East Asian and South Asian), their identification with children in these categories and their behaviour towards them. These activities were conducted as a series of games during the regular school day in the children's normal school environments, allowing us to gauge

their intergroup attitudes and behaviour in relation to naturally occurring ethnic categories in an important natural setting.

The literature on in-group bias described above addresses a joint hypothesis: that the boundaries of in-groups and out-groups coincide with the measured categories (e.g. school, ethnicity, artificially-induced group), and that individuals behave differently towards in-group and out-group members. Similarly, we would expect participants in our study to favour members of their own ethnic group if both component hypotheses are satisfied; that is, if ethnicity is salient to children's ethnic identities at this developmental stage, and if social identity shapes children's behaviour as has been demonstrated in the case of adults. If, however, this joint hypothesis is rejected, we can gain greater insight into the relationship between social identity and altruism by testing each component of it separately. Fong & Luttmer (2009), for example, find that measured identification with one's own racial group, rather than race *per se*, is a key determinant of discrimination by blacks and whites in charitable giving. We adopt this approach, and investigate whether ethnic identities coincide with ethnic categories, and whether either categories or identity predicts children's altruistic behaviour.

We begin by investigating the patterns of ethnic self-identification among our study participants. We measure the strength of children's self-identification with ethnic categories by eliciting their "perceived similarity to self" in response to photographs of children from different ethnicities. Our procedures give participants from the majority White group and the two ethnic minority groups (East and South Asians) the opportunity to identify with any or all of these ethnicities. Previous research suggests that the processes shaping the social identities of children may differ for children from majority versus minority groups. If, as proposed by social identity theory, individuals define their in-group such that it will contribute to a positive sense of self, minority children might expand their in-group to include higher status or majority ethnic categories (Blanz, Mummendey, Mielke & Klink 1998, Hornsey & Hogg 2000) This conjecture is supported by evidence showing that participants are more likely to endorse and identify with a larger superordinate category when the status of that category is experimentally manipulated to be higher (Hornsey & Hogg 2002). We use our photo selection strategy to measure children's evaluations of each of these ethnic categories with respect to sociability and competence, and interpret these as measures of status in order to inform our understanding of the patterns of ethnic identification that we observe.

We then examine participants' allocations in a dictator game in which proposers make offers to three hypothetical others, represented by photos of same-gender children from each of the three ethnic categories. We consider the observed patterns of sharing in the context of participant's identification with these groups, and directly test the hypothesis that the strength of ethnic identification is associated with children's altruistic choices.

3.2 Sample characteristics

Our study participants were recruited from public school students enrolled in Kindergarten, Grade 1 and Grade 2, in Vancouver, Canada, a highly diverse population in which Whites are the dominant group. By focusing on school-age children, we are able to gain access to a large number of subjects in a consistent institutional environment. We focus on the early grades in order to gain insight into the developmental processes shaping social identity, ethnic stereotypes and altruism among young children.

We restrict our attention to children whose own ethnicity is represented among our target photos, that is, those who were identified by their teachers as belonging to a single ethnic group that is White, East Asian (Chinese, Korean or Japanese), or South Asian (Indian, Pakistani or Sri Lankan). Of these, 214 participants are White, 186 are East Asian and only 39 are South Asian. Among the East Asian participants, over 94% were characterized by their teachers as Chinese, and the remainder as Japanese or Korean. The children in our estimation sample are drawn from 30 different Vancouver public schools. Parents of 72% of the children in participating classrooms gave consent for their child to take part in the study. With absences, the overall participation rate was 69%.

Table 3.1 reports sample characteristics by participant ethnicity. The age distributions of the White and East Asian participants are quite similar; just under one-third of subjects are five years old, slightly less than half are six years old, about one-fifth are seven years old, and a small fraction of participants are eight years old. The South Asian participants tend to be slightly older. The response rate among East Asian females was lower than other groups; as a result the proportion of East Asian subjects who are female is only 42%, compared to 52% of Whites. Over 62% of East Asians and 56% of South Asians were enrolled in English as a Second Language (ESL) programs, while only 4% of Whites were in ESL.

3.3 Experimental procedures

Teams of three to four testers were formed to work with children from each participating classroom. Each child was individually engaged in two sets of activities, which were introduced in random order as a series of games. The "sorting task" was designed to elicit participants' beliefs about different ethnic groups, and the "sharing task" or dictator game was designed to assess ethnic discrimination in children's altruistic behaviour.

3.3.1 The Sorting Task

At the beginning of each session, the researcher took a digital photograph of the child, which was immediately printed. This photo was added to a testing pack consisting of four sets of three matched photos of children who were unknown to participants (that were not used in the dictator game), two for each gender.[1]. All 13 photos (two males and two females from each of the ethnic groups, plus the child's own photograph) were shuffled and placed randomly in front of the child. The researcher asked the child to sort the 13 photographs using the following standard request format: "Pick all the children who are ___ and, leave all the children who are not ___ on the table." The child was informed that she/he was free to pick all, some or none of the 13 photographs. The photographs were shuffled and were placed randomly in front of the child before each question. In order to make sure that the child understood the nature of the task, in the first two trials, the child was asked to pick the "girls" and the "boys".

The sorting task was used to assess children's evaluations of others' competence and sociability, and their perceived similarity to others. The sociability trials required children to pick those who are nice to other children, who are happy, who have lots of friends, and who are helpful. The competence trials require children to pick those who are smart, who work hard, who read well, and who like school. The extent to which children identify with targets was assessed by asking them to "pick all the

[1] Approximately 350 head-and-shoulder photographs of five- to seven-year old White, East Asian and South Asian children were pretested for clarity of the photograph, physical attractiveness, facial expression, gender, age, and ethnicity of the child. Nine adults from four different ethnic backgrounds rated the photographs on each of these dimensions on a seven-point Likert scale. First, only photographs that received unanimous agreement on ethnicity and gender of the child were retained. These photographs were then matched on the remaining criteria (age, physical attractiveness, facial expression, and the clarity of the photograph) to create sets of three same gender children one from each of the three ethnic groups. The procedure and materials for this task were adopted from Wright & Taylor (1995)

children who are like you." For each trial, children can choose between 0 and 4 targets from each of the three ethnic groups (excluding the child's own photo). This approach allows children to associate themselves with multiple ethnic categories, and to indicate the strength of their identification by selecting more or fewer photos.

3.3.2 The Dictator Game

Like Eckel & Petrie (2011), we represent receivers in our game using photos. Boys give to boys only; girls give to girls only. All participants play the game three times, each time with 12 stickers that they can divide four ways (self, White, East Asian, South Asian). By allowing children to view the three target photos simultaneously, our intention is to increase the salience of phenotypic differences among them. At the same time, however, this approach may increase the salience of fairness.[2]

3.3.3 Supplemental data

After testing was complete, each classroom teacher was asked to fill out an information sheet that included questions about each participant's characteristics, including their ethnicity, gender, home language, and English language proficiency, and to provide aggregate information about the overall composition of the classroom (including children who did not participate in the study). Teachers' assessments of children's ethnicity may be informed by their knowledge of the child's phenotype, the parents' phenotype, and the family's home language and culture. Finally, we collected participants' residential postal codes on the Parent Permission Form required for all participants, and linked these postal codes to 2001 Census information about the characteristics of the population residing in the same Dissemination Area (DA). DAs are geographic areas designated for the collection of Census data, and are composed of one or more neighbouring blocks with a population of 400 to 700 persons. Details of the linking of postal codes to DAs are provided in the Data Appendix.

3.4 Ethnic identity

As described earlier, both theory and previous evidence suggest that children from dominant majority groups tend to identify positively with their own ethnic group,

[2] We engaged a smaller number of subjects in a "sequential" version of our procedure, in which children played a series of two-person games against the targets. This method produced very noisy responses.

while minority children may also identify with higher status ethnic majority categories. We begin by investigating the status of the three target ethnic groups among our participants, measured by their evaluations of the sociability and competence of each target ethnicity. We then evaluate the tendency of children in each ethnic category to identify with the target ethnicities according to our measure of perceived similarity.

3.4.1 Results

The first column of Table 3.2 shows the mean number of photos selected by White, East Asian and South Asian participants respectively from each target group in response to the sociability, competence and perceived similarity items. Out of a maximum of four photos per target ethnicity (two boys and two girls), participants selected between 2.3 and 2.9 photos on average in response to the items in the sociability and competence scales. The next three columns show the two-way differences (between target ethnicities) in the average number of photos that participants choose, for each group of participants. The final column reports the frequency with which participants chose their own photos.

The first panel shows results for evaluations of sociability and the second panel for competence. All three groups of participants evaluated the White targets as being both the most social and the most competent, followed by the East Asian targets and finally the South Asian targets. While their ranking of target ethnicities was the same as the other two groups, the magnitude of the distinctions between them was smallest among South Asian participants. However, none of these distinctions is statistically significant.

The third panel of Table 3.2 shows that, on average, White and East Asian participants selected 1.2 out of 4 photos from each target ethnicity as being "like them", and South Asian participants chose slightly more (1.4). These numbers are substantially lower than the average number of photos selected in response to the sociability and competence items, reflecting relatively low selection rates of opposite gender photos in response to the perceived similarity item. Overall, White participants selected White photos about 50 percent more often than East Asian photos (0.60/1.2), and about 70 percent more often than South Asian photos (0.89/1.2), and these differences are statistically significant. East Asian participants selected East Asian photos about 45 percent more often than South Asian photos (0.56/1.2); they selected East Asian photos only 16 percent more often than White photos ($-0.20/1.2$), and this latter

difference is not statistically significant. South Asian participants selected about 39 percent more South Asian photos than White photos (−0.55/1.4), and about 36 percent more South Asian photos than East Asian photos (−0.50/1.4). Although the magnitudes of these differences are substantial, they are not estimated very precisely due to the small number of South Asian participants.

3.4.2 Discussion

Research involving children from dominant majority groups has shown that from the age of three, they evaluate both their own gender and ethnic group more positively than others, like them more and feel more similar to them (e.g., Aboud 1988, Martin, Ruble & Szkrybalo 2002, Nesdale, Maass, Griffiths & Durkin 2003). Our results for White participants are weakly consistent with this evidence. White participants exhibit stereotypic beliefs and rank the three ethnic categories from Whites (highest) to South Asians (lowest) according to both sociability and competence, although these distinctions are not statistically significant. Their patterns of perceived similarity reveal a clear sense of identification with the White target, and relatively weak identification with the targets from the lower status minority ethnic groups. Whites perceive themselves to be least similar to the South Asian target, which they evaluate least favourably.

East Asian participants share Whites' rankings of the three ethnic groups according to sociability, situating their own group between Whites and South Asians. Again, however, there is no statistically significant difference between their evaluations of the three targets. This result differs from the findings of previous research that East Asian children aged five to seven tend to identify East Asian children rather than White children as being good at math (Ambady, Shih, Kim & Pittinsky n.d.). We speculate that our results may differ from these because our competence items are different. "Reads well" may be a trait associated with White children, of whom a much larger proportion are native English speakers. Reading and oral skills may be the most visible markers of "being smart" for children in this age group, if they are the most salient academic competencies sought by teachers in the early grades. None of our competency items specifically refers to quantitative skills.

Unlike Whites, East Asians do not reveal a clear sense of identification with their own ethnic category. Instead, they appear to associate themselves with a superordinate ethnic group that includes the White as well as the East Asian ethnicity, but excludes South Asians. These results are consistent with previous evidence that

minority groups may expand their in-group to include higher status or majority ethnic categories (Blanz et al. 1998, Hornsey & Hogg 2000) and that minority children express preferences for contact with the majority outgroup (e.g., Clark & Clark 1939, Clark & Clark 1947, Katz & Braly 1933, Corenblum & Annis 1993, Aboud & Doyle 1996).

Although the point estimates indicate the same ranking of Whites (highest), East Asians and South Asians (lowest), South Asians perceive smaller differences in the sociability and especially in the competence of the three target ethnicities. This pattern of results provides weak evidence that South Asians may be evaluating their own group relatively favourably, compared to prevailing stereotypes. Unlike the East Asian minority group, however, South Asians do not associate themselves with a superordinate ethnic group that includes higher status categories. Instead, like Whites, they have a clear tendency to identify more strongly with their own ethnic category than with the other two.

3.5 Altruistic behaviour

Table 3.3 characterizes the general patterns of sharing behaviour in the data. Overall, participants share on average 13.6/36 stickers or 38% of their endowment. South Asian children share fewer stickers overall (11.7) compared to Whites (13.8) and East Asians (13.7). These results are similar to the results of Gummerum, Keller, Takezawa & Mata (2008) for German children, who allocated on average between 35% and 40% of their endowment to anonymous others in a dictator game. As those authors note, these allocations are greater than both the 20% that is typically offered by adults (e.g., Camerer 2003), and the offers made by young children in two U.S. studies (Harbaugh & Krause 2000, Bettinger & Slonim 2006). Like previous authors, (e.g., Harbaugh et al. 2002, Benenson et al. 2007, Bettinger & Slonim 2006, Fehr et al. 2008), we find that children share more as they grow older; the average number of stickers shared by seven-year olds in our sample (15.9) is substantially greater than the average number shared by five-year olds (12.7). Our results also confirm previous results that girls (14.8) share more than boys (12.8) when playing the dictator game (Harbaugh et al. 2002, Gummerum et al. 2009).

The frequency distribution of the number of stickers shared by each child, presented in Figure 3.1, shows pronounced spikes at multiples of three. The modal response was nine stickers, chosen by 14.0 percent of subjects. The second highest frequency was eighteen stickers, chosen by 6.9 percent of participants. These spikes

are suggestive of non-discriminatory sharing, i.e. sharing the same number of stickers with each of the three recipients. The second column of Table 3.3 shows that 50 percent of participants chose a non-discriminatory allocation (including 5.2 percent of participants who shared zero stickers). Non-discriminatory sharing may be chosen frequently because it provides a cognitively undemanding rule of thumb (Messick 1993), or because it reflects children's developing egalitarianism or norms of fairness that may be reinforced in school environments (Fehr et al. 2008). White participants were substantially more likely (55 percent) than East Asian (45 percent) and South Asian (40 percent) participants to choose a non-discriminatory allocation, and girls were more likely (54 percent) than boys (42 percent) to do so. Non-discriminatory behaviour is slightly less frequent among five-year olds (46 percent) than among six- and seven-year olds (50 percent).

Table 3.4 presents more detailed information about sharing behaviour by participants from different ethnic categories. Columns 1-3 show that, on average, White participants shared slightly more stickers with the recipient from their own ethnic category than with the other two; East Asians participants shared slightly fewer with their own category than with the other two; and South Asian participant shared slightly fewer with the East Asian recipient than with the other two. Columns 4-6 show results for the sub-sample of participants who did not share the same number of stickers with each of the three recipients. Among participants who discriminate, these patterns of discriminatory sharing are more pronounced.

We next examine the relationship between ethnic categories, ethnic identity and altruism in the dictator game.

3.5.1 Empirical framework

In each trial t, participants choose to allocate an endowment of E stickers between themselves and one photo each of same-gender White, East Asian and South Asian children (three photos in total). We aggregate each participant's allocation to recipient photo$j = W, EA, SA$ (White, East Asian and South Asian, respectively) across all three trials to generate each participant's overall allocation to each recipient ethnicity:

$$q_{ij} = \sum_{t=1}^{3} q_{ijt}$$

Suppose that participant i has preferences over this allocation that are represented by the following utility function:

$$U_i = U\left(q_{i0}, q_{iW}, q_{iEA}, q_{iSA}, x_i; \Theta\right) \tag{3.1}$$

where q_{i0} is the number of stickers kept by the subject for themselves, q_{ij} is the number of stickers allocated to recipient photo j, x_i is a vector of individual characteristics that influence preferences (including ethnicity), and θ is a parameter vector. Subjects choose the allocation $\{q_{i0}, q_{iW}, q_{iEA}, q_{iSA}\}$ to maximize this utility function, subject to the endowment constraint. The allocations that maximize utility can be written:

$$\begin{aligned} q^*_{ij} &= f_j(x_i, \Theta) \\ q^*_{i0} &= E - (q^*_{iW}, q^*_{iEA}, q^*_{iSA}) \end{aligned} \tag{3.2}$$

We are interested in the extent to which participants' sharing allocations are systematically biased towards any ethnic categories, and particularly whether they favour their own group. We investigate two different ways of characterizing ethnic bias. In the first case, we specify our estimating equation in order to focus on the *relative* number of stickers that participants share across recipient photos, rather than the absolute number of stickers they choose to share with each recipient. Specifically, our dependent variable is an indicator for whether more than one-third of the total number of stickers shared by participant i were allocated to ethnic category j.

$$y_{ij} = \mathbb{1}_{ij} = D\colon \left[\left(q_{ij} - \frac{1}{3}\sum_j q_{ij}\right) > 0\right] \tag{3.3}$$

where $D\colon [\cdot]$ is an operator that assigns the value one when the condition inside the square brackets is satisfied, and zero otherwise. This way of defining favouritism allows participants to favour zero, one or two recipients. For example, a participant who allocates nine stickers in total could show no favouritism by choosing to allocate three stickers to each recipient (3-3-3) or by sharing no stickers at all (0-0-0), favour one of the recipients at the expense of one or both of the others (e.g. 5-3-1 or 5-2-2), or favour two recipients at the expense of a third (e.g. 5-4-0 or 4-4-1). This feature of our definition of favouritism allows for the important possibility that participants whose ethnic identities encompass two ethnic categories may favour both categories.

Our baseline model for each of these relative allocation decisions as follows:

$$y_{ij} = f\left(X_{ij}^{'}\beta + \mu_{ij}\right) \tag{3.4}$$

$$\begin{aligned} X_{ij}^{'}\beta = &\left(\beta_0 + \beta_1 \mathbb{1}_{ij}^{\text{TgtOwn}}\right) + \left(\beta_2 + \beta_3 \mathbb{1}_{ij}^{\text{TgtOwn}}\right) \cdot \mathbb{1}_i^{EA} \\ &+ \left(\beta_4 + \beta_5 \mathbb{1}_{ij}^{\text{TgtOwn}}\right) \cdot \mathbb{1}_i^{SA} \end{aligned} \tag{3.5}$$

where $\mathbb{1}_{ij}^{\text{TgtOwn}}$ indicates that the recipient j is from the same ethnic group as the participant i, $\mathbb{1}_i^{EA}$ and $\mathbb{1}_i^{SA}$ indicate that the participant is East Asian or South Asian, respectively, and β_0 to β_5 are parameters to be estimated. We assume the following specification for the stochastic error term μ_{ij}:

$$\mu_{ij} = \delta_i + \varepsilon_{ij}$$

Here δ_i is a random person effect and ε_{ij} is an idiosyncratic error term. We estimate equation 3.4 using a random effects (also known as a mixed effects) probit model.

The probit model affords a direct test of the hypothesis that participants favour their in-group. However, it does not allow us to assess the intensity of any in-group bias. In order to do so, we estimate an alternative model in which the dependent variable is the number of stickers shared with each recipient:

$$q_{ij} = g\left(X_{ij}^{'}\Theta + \varphi_{ij}\right) \tag{3.6}$$

where $X_{ij}^{'}\Theta$ takes the same form as $X_{ij}^{'}\beta$ in 3.5, and the stochastic error term φ_{ij} includes both a random person effect and an idiosyncratic error term. We estimate equation 3.6 using a negative binomial model to accommodate the count nature of the dependent variable, and cluster the standard errors at the participant level.

3.5.2 Results

Baseline results

The unit of observation is a participant/recipient pair, so the estimation sample consists of three observations for each participant, each corresponding to a different recipient. All recipients are the same gender as the participant. We begin by reporting our results from the probit model in which the dependent variable is an indicator for whether the participant favoured (gave more than one-third of the total number of stickers they shared to) a given recipient. The dependent variable takes on the value

zero in all three observations corresponding to a given participant if the child shares equally across all three recipients. If the child does not share equally across all three recipients, the dependent variable may take on the value in at most two cases, since the proportion of stickers shared with each recipient must sum to one within each set of three child/recipient observations.

The estimated marginal effects reported in the first column of Table 3.5 correspond to a specification that includes only an indicator for whether the child belongs to the same ethnic category as the recipient photo. Participants give a recipient more than a third of the total number of stickers they share in about 18 percent of cases. On average, however, they are no more likely to favour the recipient from their own ethnic category than to favour a recipient from one of the other ethnic categories.

The specification in the second column relaxes the constraint that participants from different ethnic categories behave similarly. In this specification, the omitted group is White children giving to Whites, and the tendency to favouritism is constrained to be the same across outgroups (e.g. Whites equally likely to favour East Asians and South Asians). The proportion of children who favour their own category is 26 percent among Whites (*Constant + Owngroup*), 26 percent among East Asians (*Constant + East Asian Participant + Owngroup + Owngroup x East Asian Participant*), and 23 percent among South Asians (*Constant + South Asian Participant + Owngroup + Owngroup x South Asian Participant*). The difference in the behaviour of South Asians participants and the other two groups is not statistically significant.

Only 16 percent of White children (*Constant)*, and 21 percent of South Asian children (*Constant + South Asian Participant*) favour each of the other groups, compared to 31 percent of East Asians (*Constant + East Asian Participant).* We cannot reject the null hypothesis that the behaviour of Whites and South Asians towards out-group categories is the same. However, the behaviour of East Asian children towards out-group categories is statistically significantly different from the behaviour of both Whites and South Asians.

Having established that there are significant differences between the behaviour of Whites and South Asians on one hand and East Asians on the other, we now seek to characterize the behaviour of each group in terms of in-group bias. The number of Whites who favour the White recipient is 10 percentage points larger than the number of Whites who favour each of the other recipients (*Owngroup*), and this difference is statistically significant ($p = 0.06$). In contrast, the number of East Asian participants who favour the East Asian recipient is five percentage points *lower* than the number who favour each of the other two (*Owngroup + Owngroup x East Asian Participant*),

and this difference is also statistically significant ($p = 0.03$). Note again that the difference between the behaviour of East Asian children and White children towards their own group relative to other groups comes primarily from differences in their behaviour towards the other groups; their behaviour towards their own group does not differ in any significant way. While the absolute magnitudes of these differences in favouritism are fairly small, they indicate clearly that the ethnicity of the recipient is salient to White and East Asian children's sharing behaviour in the classroom environment.

According to the point estimates, the number of South Asians who favour the South Asian recipient is 8 percentage points larger than the number of South Asians who favour each of the other two (*Owngroup + Owngroup x South Asian Participant*). The estimate of in-group bias is very imprecise, however, and is not statistically significant ($p = .51$). Given the small South Asian sample size and the resulting problem with precision, we restrict our attention to in-group bias among Whites and East Asians and exclude South Asians from our estimation samples in the remaining regressions.

The specification reported in the third column of Table 3.5 excludes South Asians from the estimation sample and includes a variable indicating that the recipient is South Asian and its interaction with an indicator that the participant is East Asian. Its' purpose is to provide a basis for testing the constraint that White and East Asian children do not distinguish between out-groups. The results show that White children favour the East Asian (*Constant*) and South Asian recipients (*Constant + South Asian Target*) with equal frequency, and East Asian children favour the White (*Constant + East Asian Participant*) and South Asian recipients (*Constant + East Asian Participant + South Aisna Target + South Asian Target x East Asian Participant*) with equal frequency. In other words, neither ethnic group distinguishes between the two out-group categories.

We next report results for the model where the dependent variable is the number of stickers shared with each recipient. This alternative specification frames the central hypothesis in terms of levels, rather than indicators of favouritism, in order to provide a sense of the intensity of bias. The estimated marginal effects in column 5 show that White subjects share slightly less than one-third of a sticker more on average with the White target than with either of the other two recipients (*Owngroup*). However, East Asian participants show no in-group bias; if anything, they share fewer stickers with their own group than with other groups (*Owngroup + Owngroup x East Asian Participant*). This difference in patterns of in-group bias between Whites and East

Asians is statistically significant. As before, the relevant point estimates indicate that South Asians may show somewhat less in-group bias than Whites in terms of the number of stickers shared (*Owngroup + Owngroup x South Asian Participant*), but this difference is not statistically significant. The results in column 6 provide no evidence that Whites and East Asians distinguish between the two out-groups when sharing stickers.

Specifications with demographic interactions

We next investigate whether differences in the gender and age distributions of White and East Asian participants may explain any of the observed differences between their sharing behaviour. The specification reported in the second column of Table 3.6 includes an indicator that the participant is female, interacted with all the variables in our main specification (reproduced in column 1). In the third column, we add the participant's age in years, again interacted with all of the variables in our main specification. The inclusion of these variables does not alter the main results, although larger standard errors mean that the in-group bias effect among White is no longer statistically significant. The results reveal no statistically significant differences in in-group bias according to age or gender. We see weak (statistically insignificant) evidence that the absence of in-group bias among East Asians is driven primarily by the behaviour of boys.

When we consider the intensity of in-group bias, measured in terms of the number of stickers shared, the differences by gender among East Asian children are marginally statistically significant. The results in column 6, which correspond to a specification that includes interactions for both age and gender, show that both White boys ($p = 0.05$) and White girls ($p = 0.03$) share more stickers with their own group than with other groups on average. East Asian boys show out-group bias; they share fewer stickers on average with their own group than with other groups ($p = 0.03$). East Asian girls demonstrate no in-group or out-group bias with respect to the total number of stickers shared ($p = 0.76$). The difference between White and East Asian boys in their behaviour towards their own group is statistically significant ($p = .004$), but the difference between White and East Asian girls is not ($p = .52$).

Robustness checks

We next investigate the possibility that children's behaviour may be influenced by the ethnicity of the tester who administered the experimental procedures. Over 58 per-

cent of participants interacted with a White/European tester, 12 percent with an East Asian tester, 6 percent with a South Asian tester, and 25 percent with a Hispanic, Middle Eastern or mixed White/Korean tester. The specifications reported in the columns 2 and 5 of Table 3.7 include an indicator that the tester is non-white, interacted with all of the variables in our main specifications (reproduced in columns 1 and 4 of Table 3.6). The results show that tester ethnicity effects have no important effect on measured in-group bias. Interestingly, we see weak evidence that East Asian participants may share more stickers on average when they interact with a non-White tester.

Differences in socioeconomic status between White and East Asian subjects are also a potentially confounding influence. Table 3.8 reports the average, for each of our ethnic groups, of four neighbourhood (DA)-level variables: the proportion of household heads who immigrated to Canada in the previous five years, the proportion whose education level is high school completion or less, the proportion whose incomes are below the low-income cutoff defined by Statistics Canada, and mean family income from all sources. Among our participants, Whites on average are drawn from relatively high socioeconomic status neighbourhoods, with the lowest immigrant density, the lowest poverty rate, the fewest household heads who had not gone beyond high school, and highest mean family income.

We create a variable indicating whether a participant lived in a neighbourhood that was in the bottom half of the distribution of neighbourhood poverty rates among the families in our sample. The specifications reported in columns 3 and 6 of Table 3.7 allow all the coefficients in our main specification to differ for students living in neighbourhoods in the top and bottom halves of this distribution. The results provide no evidence that differences in neighbourhood income explain any of the observed differences in the behaviour of White and East Asian participants.

Specifications with perceived similarity

We next explore the direct effects of children's identification with ethnic categories on their sharing behaviour. The key explanatory variable in the specifications reported in Table 3.9 is a measure of the participant's identification with the ethnic category. In the favouritism model, we define this indicator with respect to the number of photos chosen of a given target ethnicity in response to the perceived similarity question, relative to the average number of photos selected by the participant in response to this question. The indicator takes on the value one if more than one-third of the total number of target photos selected by participant i pertains to ethnic category

j. Like the dependent variable, this indicator can take on the value one either never, once or twice for each participant, allowing for the possibility of ethnic identities that incorporate up to two ethnic categories. In the count model of stickers shared, we measure ethnic identification as the number of target photos selected from the ethnic category in response to the perceived similarity question.

The results in column 1 of Table 3.9 show that our indicator of ethnic identification alone explains none of the variation across participants and recipients in favouritism as measured by our indicator variable. When we relax the constraint that the sharing behaviour of White and East Asian children is the same in column 2, we see that White children more frequently favour recipient ethnicities if they identify with them, but East Asian children do not ($p = 0.45$). This difference between White and East Asian children in the association of ethnic identity and in-group bias in sharing behaviour is statistically significant ($p = 0.04$).

The results in columns 4 and 5 correspond to the number of stickers shared, where ethnic identification is likewise defined as a count variable (number of photos selected in response to perceived similarity question). In this case, we see in column 4 a statistically significant relationship between the intensity of ethnic identification and the number of stickers shared in the pooled sample. In other words, the average participant in the sample shares more stickers with recipients from ethnic categories that they identify more strongly with. Column 5 shows that this behaviour is strong and statistically significant among Whites, and weaker and statistically insignificant among East Asians ($p = 0.47$).

3.6 Conclusion

Our results confirm previous research that demonstrates that young children, like adults, engage in altruistic behaviour and show a strong tendency towards egalitarianism. We gain further insight into sharing behaviour by examining whether children show a preference for their social in-group when they deviate from an equal allocation of stickers across the three targets. We find clear evidence that participants from the majority White category behave more altruistically towards a White recipient than either an East Asian or South Asian recipient. Together with strong, independent evidence that White children identify with the White ethnic category, this result is consistent with the predictions of social identity theory. The statistically significant relationship between the indicator of ethnic identity and the indicator of favouritism provides direct support for the hypothesis that ethnic identity influences

altruism among White children. While they evaluate the South Asians somewhat less favourably than East Asians in terms of sociability and competence, White children do not identify more with one out-group than the other, and do not favour one out-group more frequently than the other when allocating stickers in the dictator game. These results are consistent with Brewer's (1999) conclusion that in-group attachment is psychologically primary, and attitudes towards out-groups are not.

The estimated differences across recipients in the number of stickers shared by White participants is quantitatively small; on average they shared less than one-third of a sticker more with the White recipient compared to the two out-group recipients. Whether this result indicates that ethnic identity is an important or unimportant determinant of children's behaviour is a matter of interpretation. While small, the fact that an effect is detectable in an experimental set-up where children's sense of egalitarianism may be primed – both by being shown the three photos simultaneously and by being assessed in the context of school social environments that typically emphasize fairness – could be viewed as evidence that ethnic identity is a powerful factor shaping altruistic preferences among White children.

The results for East Asian children exhibit a different pattern. Unlike Whites, East Asian's social identities extend beyond their own ethnic category to include the majority White category, while still excluding the South Asian category. East Asian children's sharing behaviour also does not conform to the predictions of social identity theory. Although they show substantially less egalitarianism than White participants, we find no evidence of in-group bias in the dictator game among East Asian children, and no evidence that perceived similarity plays any role in their sharing decisions. Instead, East Asian participants show bias in favour of the White and South Asian recipients relative to the East Asian recipient; East Asians are as likely as Whites to favour their own category, but twice as likely to favour each of the other categories. Again, the importance of this result, given the magnitude of the estimated effect, is a matter of interpretation.

The factors underlying these differences between children from different ethnic backgrounds are unclear. It is tempting to interpret the patterns of perceived similarity among East Asians as evidence that they are defining their social identities to include higher status or majority ethnic categories. However, the sorting task results do not reveal any stark differences in children's evaluations of the sociability or competence of Whites and East Asians that would suggest clear differences in perceived social status. Moreover, while South Asians are the lowest status of the three ethnic

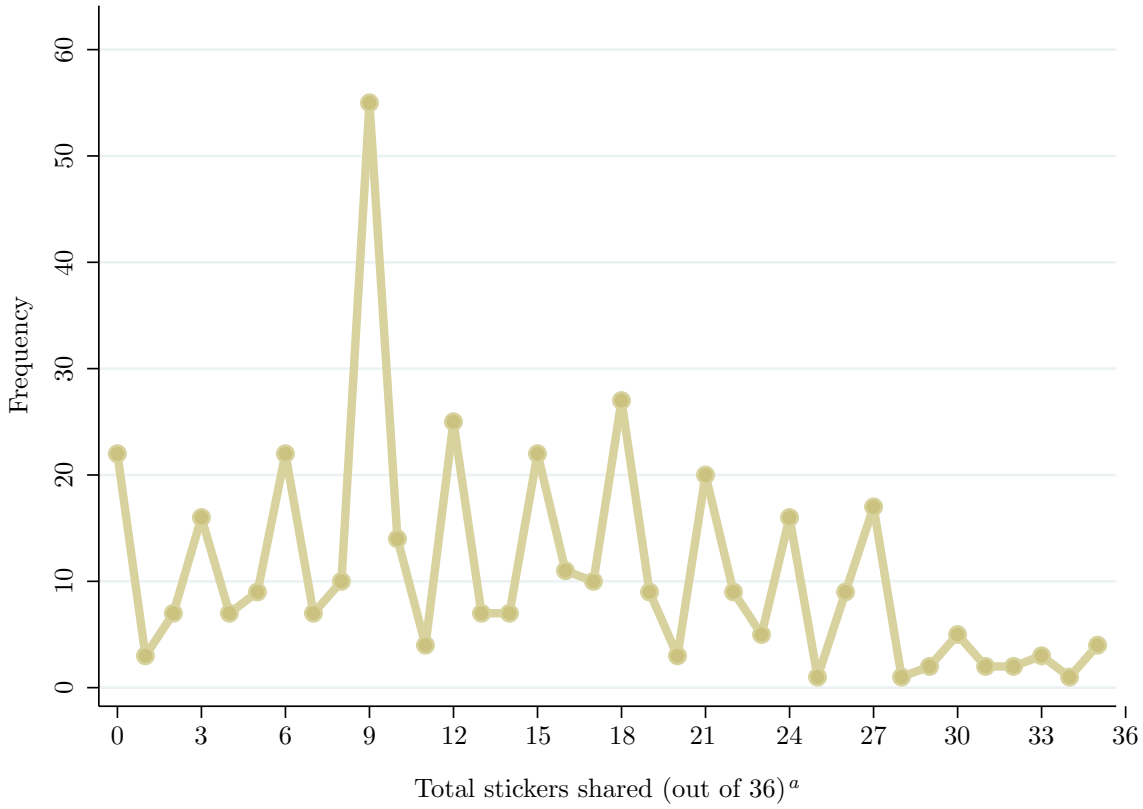
categories according to the sorting task evaluations, there is no evidence that South Asian participants expand their social identities to include higher status categories.

East Asian children's favoritism towards other ethnic categories is also puzzling. East Asian cultures have been characterized as "collectivist" (Hofstede 1980) relative to more individually oriented North American cultures. Previous research has found that while participants from collective cultures may demonstrate strong in-group bias with respect to naturally occurring groups, they tend not to show in-group bias towards artificially induced group associations (e.g., Triandis 1995, Buchan et al. 2006). Whether the weak evidence of favoritism towards anonymous others from different ethnic categories among our East Asian participants reflects a social norm in East Asian communities remains a question for future research.

Regardless of their underlying determinants, these clear differences in patterns of discriminatory sharing between White and East Asian participants imply that care is required when interpreting responses in experimental games played with children from a variety of ethnic backgrounds. For example, Goeree et al.'s (2010) finding that dictator offers of ten to twelve-year-olds were no more generous when the proposer and the recipient shared the same race mirrors the results from our pooled sample (including participants from all three ethnic groups). In our study, the pooled analysis is misleading: it masks significant in-group bias among Whites and out-groups bias among East Asians. Goeree et al.'s (2010) sample is composed of girls of whom 51 percent are Caucasian, 27 percent Asian, and 22 percent from other and mixed-race groups. Further analysis of these and other data by disaggregated ethnic groups would be interesting in light of our findings.

Figures

Figure 3.1: Frequency of total number of stickers shared



[a] Total number of stickers shared by participant i is the sum across all three trials t, and all three target photos $j = W, EA, SA : \sum_j \sum_{t=1}^{3} q_{ijt}$.

Tables

Table 3.1: Sample characteristics, by participant ethnicity

		Whites		**East Asian**		**South Asian**	
		Frequency	Percent	Frequency	Percent	Frequency	Percent
Age	five	64	29.9	59	31.7	10	25.6
	six	105	49.1	79	42.5	14	35.9
	seven	43	20.1	39	21.0	13	33.3
	eight	2	0.9	9	4.8	2	5.1
Female		111	51.9	78	41.9	19	48.7
ESL		9	4.2	115	61.8	22	56.4
Observations		214	100.0	186	100.0	39	100.0

Table 3.2: Evaluations of sociability and competence and perceived similarity to ethnic phenotypes, by participant ethnicity

	Mean chosen per target ethnicity, excluding self		Difference in means across target ethnicities, excluding self[a]			
Photos chosen by	Mean	(st. dev.)[b]	White-East Asian	White-South Asian	East-South Asian	**N**
Sociability						
Whites	2.94	(0.89)	0.22	0.44	0.22	281
East Asians	2.31	(1.07)	0.20	0.42	0.22	253
South Asians	2.64	(1.11)	0.15	0.31	0.16	60
Competence						
Whites	2.59	(1.04)	0.20	0.29	0.10	280
East Asians	2.56	(1.00)	0.14	0.45	0.32	255
South Asians	2.39	(1.11)	0.06	0.18	0.12	62
Perceived similarity						
Whites	1.21	(1.25)	0.60^{***}	0.89^{***}	0.28^{*}	280
East Asians	1.21	(1.25)	-0.20	0.36^{***}	0.56^{***}	253
South Asians	1.43	(1.30)	0.04	-0.55^{*}	-0.50	60

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[a] Wilcoxon Rank Sum tests for differences in means.

[b] Range = 0-4 (two photos of boys and two photos of girls, from a given target ethnicity).

Table 3.3: Characteristics of sharing behaviour, by age and gender

		Average number shared	Frequency not discriminating[a]
All		13.6	0.50
White		13.8	0.55
East Asian		13.7	0.45
South Asian		11.7	0.40
Age	five	12.4	0.46
	six	13.3	0.50
	seven	15.2	0.50
Female		14.8	0.54
Male		12.8	0.42

[a] Proportion of participants who allocate the same number of stickers to each of the three targets.

Table 3.4: Average number of stickers shared

	All participants			**Participants who discriminate[a] among targets**		
	Shared with:			**Shared with:**		
Shared by:	White	East Asian	South Asian	White	East Asian	South Asian
Whites	4.8	4.5	4.5	5.3	4.7	4.8
East Asians	4.7	4.5	4.6	5.1	4.8	5.0
South Asians	4.1	3.9	4.1	3.7	3.1	3.9

[a] Participants who do not allocate the same number of stickers to each of the three targets.

Table 3.5: Results, sharing in the dictator game

	Favoritism (probit)[a]			**Number shared (poisson)**[b]		
Independent variables[c]	(1)	(2)	(3)[d]	(4)	(5)	(6)[d]
Owngroup[e]	−0.00	0.10**	0.07*	0.09	0.28**	0.30*
	(0.02)	(0.05)	(0.04)	(0.08)	(0.12)	(0.12)
East Asian Participant		0.15***	0.11***		0.11	0.18
		(0.04)	(0.03)		(0.29)	(0.30)
Owngroup x East Asian Participant		−0.15***	−0.12***		−0.39**	−0.46**
		(0.04)	(0.04)		(0.17)	(0.17)
South Asian Participant		0.05			−0.57	
		(0.04)			(0.46)	
Owngroup x South Asian Participant		−0.08			−0.10	
		(0.05)			(0.17)	
South Asian Target			0.01			0.04
			(0.06)			(0.09)
South Asian Target x East Asian Participant			−0.01			−0.13
			(0.09)			(0.16)
Constant	0.18***	0.16***	0.18***			
	(0.02)	(0.03)	(0.02)			
Mean predicted value				4.54	4.54	4.60
Number of observations	1294	1294	1183	1294	1294	1183
Number of participants	434	434	396	434	434	396

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

[a] Dependent variable is an indicator that participant i gave more than one-third of the total number of stickers she shared to target j. Random person effects accounted for in estimation.

[b] Dependent variable is the number of stickers shared by participant i with target j. Standard errors clustered at the participant level.

[c] Reported results are marginal effects (dy/dx). For indicator variables, marginal effects are those associated with a discrete change in value from 0 to 1.

[d] Estimation sample includes Whites and East Asians only.

[e] Participant ethnicity equals Target ethnicity

Table 3.6: Results, sharing in the dictator game, with demographic variables, White and East Asian participants only

	Favoritism (probit)[a]			**Number shared (poisson)**[b]		
Independent variables[c]	(1)	(2)	(3)	(4)	(5)	(6)
Owngroup[d]	0.07*	0.08	0.07	0.28**	0.35	0.55*
	(0.04)	(0.05)	(0.07)	(0.12)	(0.25)	
East Asian Participant	0.11***	0.12***	0.09	0.11	0.45	0.12
	(0.03)	(0.04)	(0.06)	(0.29)	(0.43)	(0.58)
Owngroup x East Asian Participant	−0.13***	−0.17***	−0.19***	−0.40**	−0.68***	−0.92***
	(0.04)	(0.04)	(0.05)	(0.17)	(0.23)	(0.30)
Female		−0.06	−0.06		0.70*	0.73*
		(0.04)	(0.04)		(0.39)	(0.39)
Owngroup x Female		−0.02	−0.02		−0.13	−0.15
		(0.07)	(0.07)		(0.24)	(0.24)
Female x East Asian Participant		−0.04	−0.04		−0.59	−0.72
		(0.06)	(0.06)		(0.54)	(0.53)
Owngroup x Female x		0.19	0.18		0.72*	0.77*
East Asian Participant		(0.14)	(0.14)		(0.43)	(0.45)
Age			−0.00			0.36
			(0.03)			(0.25)
Owngroup x Age			0.01			−0.19
			(0.05)			(0.13)
Age x East Asian Participant			0.03			0.33
			(0.04)			(0.35)
Owngroup x Age x			0.04			0.26
East Asian Participant			(0.07)			(0.21)
Constant	0.18***	0.21***	0.21***			
	(0.02)	(0.03)	(0.03)			
Mean predicted value				4.59	4.59	4.57
Number of observations	1183	1183	1183	1183	1183	1183
Number of participants	396	396	396	396	396	396

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

[a] Dependent variable is an indicator that participant i gave more than one-third of the total number of stickers she shared to target j. Random person effects accounted for in estimation.

[b] Dependent variable is the number of stickers shared by participant i with target j. Standard errors clustered at the participant level.

[c] Reported results are marginal effects (dy/dx). For indicator variables, marginal effects are those associated with a discrete change in value from 0 to 1.

[d] Participant ethnicity equals Target ethnicity

Table 3.7: Results, sharing in the dictator game with tester ethnicity and neighbourhood income, White and East Asian participants only

	Favoritism (probit)[a]			**Number shared (poisson)**[b]		
Independent variables[c]	(1)	(2)	(3)	(4)	(5)	(6)
Owngroup[d]	0.07*	0.08	0.09**	0.28**	0.25	0.30*
	(0.04)	(0.05)	(0.05)	(0.12)	(0.16)	(0.16)
East Asian Participant	0.11***	0.10**	0.10**	0.11	0.35**	0.44
	(0.03)	(0.04)	(0.05)	(0.29)	(0.15)	(0.47)
Owngroup x East Asian Participant	−0.13***	−0.12**	−0.19***	−0.40**	−0.47**	−0.54***
	(0.04)	(0.05)	(0.04)	(0.17)	(0.24)	(0.20)
Non-White Tester		0.05			0.17	
		(0.05)			(0.16)	
Non-White Tester x		0.02			0.07	
East Asian Participant		(0.07)			(0.22)	
Owngroup x Non-White Tester		−0.02			−0.06	
		(0.07)			(0.25)	
Owngroup x Non-White Tester		−0.02			−0.08	
East Asian Participant		(0.10)			(0.36)	
Low Income			−0.01			0.27
			(0.05)			(0.44)
Low Income x East Asian Participant			0.01			−0.41
			(0.07)			(0.61)
Owngroup x Low Income			−0.05			0.05
			(0.07)			(0.26)
Owngroup x Low Income x			0.20			0.18
East Asian Participant			(0.15)			(0.37)
Constant	0.18***	0.16***	0.19***			
	(0.02)	(0.03)	(0.03)			
Mean predicted value				4.59	4.59	4.63
Number of observations	1183	1183	1114	1183	1183	1183
Number of participants	396	396	373	396	396	373

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

[a] Dependent variable is an indicator that participant i gave more than one-third of the total number of stickers she shared to target j. Random person effects accounted for in estimation.

[b] Dependent variable is the number of stickers shared by participant i with target j. Standard errors clustered at the participant level.

[c] Reported results are marginal effects (dy/dx). For indicator variables, marginal effects are those associated with a discrete change in value from 0 to 1.

[d] Participant ethnicity equals Target ethnicity

Table 3.8: Census neighbourhood characteristics, by participant ethnicity

Characteristic	White	East Asian
Immigrants (%)	38.6	56.0
Poverty (%)	15.2	23.0
High school or less (%)	23.7	38.0
Mean family income (CAD/year)	105,128	78,034
Observations	214	186

Table 3.9: Results, sharing in the dictator game, with indicator of ethnic identification, White and East Asian participants

	Favoritism (probit)[a]			**Number shared (neg. binomial)**[b]		
Independent variables[c]	(1)	(2)	(3)	(4)	(5)	(6)
Identifies with Target Ethnicity	0.02	0.08*		0.24**	0.35***	
	(0.03)	(0.04)		(0.10)	(0.13)	
East Asian Participant		0.10***	0.11**		0.37	0.11
		(0.03)	(0.03)		(0.41)	(0.29)
Identifies w ith Target Ethnicity x		−0.09**			−0.24	
East Asian Participant		(0.04)			(0.20)	
Owngroup			0.07*			0.28**
			(0.04)			(0.12)
Owngroup x East Asian Participant			−0.13***			−0.40**
			(0.04)			(0.17)
Constant	0.22***	0.18***	0.18***			
	(0.02)	(0.02)	(0.02)			
Mean predicted value				4.60	4.59	4.59
Number of observations	1089	1089	1183	1089	1089	1183
Number of participants	363	363	396	363	363	396

$^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$

[a] Dependent variable is an indicator that participant i gave more than one-third of the total number of stickers she shared to target j. "Identifies with target ethnicity"defined as indicator that more than one-third of photos selected in response to perceived similarity question were from ethnic group j. Random person effects accounted for in estimation.

[b] Dependent variable is the number of stickers shared by participant i with target j. "Identifies with target ethnicity"defined as number of photos of ethnic group j in response to perceived similarity question. Standard errors clustered at the participant level.

[c] Reported results are marginal effects (dy/dx). For indicator variables, marginal effects are those associated with a discrete change in value from 0 to 1.

Chapter 4

The Evolution of Moral Punishment in Small Groups

4.1 Introduction

The dominant paradigm in contemporary economics is the idea of an independent agent maximizing her welfare subject to constraints. If agents seem to behave in contradiction to this paradigm, this is usually considered a puzzle and attributed to a lack of understanding of the constraints the agent faces or a misspecification of the agent's preferences in a given model.

Experimental economists and psychologists try to avoid the influence of unobserved constraints or preferences by placing individuals into stylized situations in which all aspects of the decision making problem are completely transparent, including all payoffs. But even in the simplest situations like the ultimatum game, norms seem to play a guiding role: Subjects show a concern for fairness and expect others to be fair as well. People who offer 'unfair' amounts in the ultimatum game are regularly punished by the responder, even if punishment is costly. Many responders reject a low offer, eliminating all surplus for both agents.

Unfair or, more generally, immoral behaviour is punished in many real world situations, too. Why is this? A purely functional explanation is certainly insufficient. It is easy to see how the threat of punishment can enforce moral behaviour, but the experimental evidence points to a deeper issue. Why would anyone punish immoral behaviour, even if it is costly for the punisher, does not generate a direct personal benefit and happens in a one-time interaction like most ultimatum games? If we happen to see a rioter destroying a shop window, we have an urge to punish him,

even if we have neither a relation to the shop owner nor to the perpetrator. And even if punishment of defection was functional in establishing cooperation, why would anyone bear a *private* cost of defending a general norm of cooperative behaviour, a public good.

This chapter investigates a possible way for morality and moral punishment to evolve. But what exactly is moral punishment? Moral punishment, as discussed in this chapter, has two defining features:

- Moral punishment is triggered by an observed *action* that violates a *norm*, regardless if the action harms the observer in any way or not.[1]
- The observation of immoral actions leads to strong negative feelings in the observer that can only be compensated by *personally* exerting effort (incurring a cost) to punish the offender, even if there is no personal gain to be expected by the punisher. The punisher is *determined* to punish as soon as she/he observes immoral behaviour.

Two additional features of moral punishment characterize it as a social phenomenon:

- First, moral behaviour is usually not enforced by one individual alone. In general, it needs the threat of punishment by more than a single punisher to change the equilibrium behaviour of players in a prisoners dilemma type of situation.[2]
- Second, the individual cost of punishing immoral behaviour decrease if there is a group of punishers that act collectively. A man in a group of five or six others punishing another individual runs a much smaller risk of serious injuries than in a one-on-one fight.

In today's world, moral punishment can be dysfunctional from an individual perspective. If the interaction is a one-time event, punishing a perpetrator seems to be a waste of resources for the punisher.

So why do we punish anyhow? This chapter argues that human behaviour could be genetically hard coded in this particular dimension. People follow instincts even if they are indeed dysfunctional today. For the evolutionary argument to be persuasive,

[1] Thus, it is more than just punishing a perpetrator as retaliation for harm done to the punisher himself or a threat to prevent a perpetrator from doing any harm in the first place.

[2] The case of a single mutant who is able to enforce behavioural change will be dicussed as well; but it does not need to be assumed for moral punishment to emerge.

however, it must be explained, why moral punishment was advantageous for reproduction at the times when human behaviour was still exclusively shaped by natural selection. Showing that the genuine circumstances of human evolution allowed for the co-evolution of morality and morally motivated punishment is the main contribution of this research.

What do we know about the circumstances of early human evolution? More than 99% of human evolution took place in small groups. About 6-5 million years ago the human line of great apes split off of what should develop into our closest relatives, the Chimpanzees (Pan troglodytes) and Bonobos (Pan paniscus). At this time our earliest ancestors left the tropical rain forest and started to adapt to life in open savanna woodlands. About 2 million years ago, some of the woodland apes started to use stone tools and included much more meat in their diet. Subsequently, their brain size expanded towards modern human size until about 200'000 years ago Homo sapiens emerged (Peterson & Wrangham 1997, p. 60-62).

Throughout this time, the human ancestors lived in small foraging bands. Only with the neolithic revolution about 11'000 years ago, foraging was replaced by agriculture and a sedentary life style, which gave rise to increased population density and more complex social entities like clans, chiefdoms and, finally, states.

In order to evaluate if moral punishment could have its origins in human evolution, a model needs to reflect important stylized facts about the environment that human ancestors lived in. They capture the situation in the Early and Middle Pleistocene, before the development of language made cultural transmission of norms possible. The closest proxy for social behaviour of Homo habilis and Homo erectus, the human ancestors in this time, can be found by observing chimpanzees. The model shows that the following stylized facts about the environment of early human evolution were sufficient for the development of morality (e.g. Peterson & Wrangham 1997, Stanford 2001, in particular p. 52-59 and table 3.2 on p. 71):

- Small groups - 4-20 adult or adolescent males. Maximum group size was restricted by cognitive capabilities of human ancestors[3] and the necessity to sustain the group while travelling and hunting together. Actual group size was mainly determined by the quality of a group's habitat

[3] Based on the correlation between group size and the size of the neocortex in modern monkeys and apes, Dunbar (1992) estimates the *maximum* total group size for chimpanzees to 60 and for Homo habilis and Homo erectus to 80 and 110, respectively. Just looking at those males competing for reproduction, leaving out females, infants and old individuals gets to the range suggested here.

- Group members randomly recombined into travelling or hunting parties on a very frequent basis (fission-fusion system). The recombinations happened multiple times a day and did not follow any particular order
- Strong territoriality among males with infrequent, but often lethal inter-group contact
- Repeated changes to the quality of the environment, especially when human ancestors were leaving their stable habitats in the tropical rain forest and spread around the globe

Group size is measured as the number of adult or adolescent males because the model presented in this chapter assumes that only males are actively competing for reproductive success. The role of females is thought to be rewarding reproductive payoff in the form of offspring to the most successful male(s). Females might judge the behaviour of other group members on moral grounds as well. I assume, however, that their selection of mates only depends on the pay-off (e.g., food) that each individual male is able to achieve. The behaviour of infants and the group's old is ignored.

Competition between males is modelled as a prisoner's dilemma game. Group members are continuously selected into random pairs to play the game. All members of the small group observe what actions are played in the prisoners dilemma and may punish actions that they consider immoral. The model shows that only individuals with a genetic determination to punish will do so. Moreover, only punishers of defection will be able to enforce cooperation and sustain *large* groups that are stable against violent inter-group contact.

However, the *initial* mutant who is genetically determined to punish can only prosper in very *small* groups. Therefore, in the model presented here, punishment of defection cannot evolve in stable environments with groups of constant size. Only fluctuations in group size allow for both the initial mutant to succeed and for cooperation and punishment of defection to co-evolve and spread in the population.

The analysis proceeds in three steps. First, the formal model environment is introduced and the various assumptions are discussed in detail. The second part of the chapter analyzes behaviour and payoffs in *one* period. It is shown that none of the players voluntarily considers any of the actions in the game immoral and, therefore, no actual punishment of any action occurs unless there is genetic determination to punish. Then the first main result of this research is reported: Genetically determined punishers have a *relative* advantage over non-punishers if the group is sufficiently small.

Third, the medium and long term dynamics of the evolution of moral punishment in a small group are discussed. Ultra-long dynamics of the underlying evolutionary process involving very unlikely events like simultaneous mutations of a group of agents are not considered in this chapter.[4]

The second main result of the chapter is established: Cooperation evolves as the norm shared by all players of a population. All players will finally become genetically determined punishers of defection. The dynamics leading to this result are discussed step by step:

First, an arbitrary mutant is introduced into a group of non-determined players. He holds a non-empty moral code, a set of actions he considers immoral. It is shown that, for a sufficiently small group, the mutants grow in number and fix their type in the group, so that all group members have the same non-empty moral code.

Evolutionary stability against further mutations is discussed for different group sizes. It is shown that in very small groups only players who punish all the time are evolutionary stable. The reason is that, in a very small group, the relative advantage of a punisher is the bigger, the more actions he punishes. The model presented here implies, however, that *big* homogeneous groups are evolutionary stable, irrespective of the type of their members.

If the mutants punish defection, all group members change their behaviour and cooperate when the number of punishers becomes big enough. Throughout the transition period to cooperation both non-punishers and punishers defect. Moreover, those players who punish defection will start to cooperate *last*. Thus, if moral behaviour is established as a consequence of active punishment, it is necessarily accompanied by hypocrisy.

The final part of the analysis shows that a volatile environment is needed for a group of genetically determined punishers to evolve. Moreover, in the presence of deadly intergroup contact, only punishers of defection can sustain against all other types and finally fix their type in the entire population. Section 4.5 concludes.

[4] For a discussion of timing in evolutionary models and the terminology used here see Samuelson (1997, p.23-29) and Binmore & Samuelson (1994)

4.2 The model

This research proposes a "small group, continuous interaction model" of human evolution: There is a group of agents of finite size n_t.

$$n_t \leq \min\{n_{max}, N_{max}\} \tag{4.1}$$

Group size must neither exceed a theoretical maximum n_{max} nor the carrying capacity of the group's habitat N_{max}. The parameter n_{max} reflects the constraints put on group size by cognitive capacities and the need to feed all individuals simultaneously while moving as a group. It is assumed to be constant over time.[5] N_{max} may vary, reflecting changes in the environment. Time is infinite, and divided into periods indexed by $t = 0, 1, 2, \ldots$.

The game Γ^e: In every period, all group members repeatedly play a two-stage game Γ^e. The game is played infinitely often within a given period, so that average payoffs over that period equal their expected values.

In the first stage of Γ^e, nature randomly selects two players to play a 2x2 normal form game Γ - the prisoners dilemma with payoff values y shown in figure 4.1.[6] Each one of the two selected players can either cooperate (C) or defect (D). Let the actions in Γ be denoted by $s \in S \equiv \{C, D\}$. *Every* player in the group observes the actions of the two players in Γ with certainty.[7]

In the second stage of Γ^e *every* member of the group has the option of punishing either one or both of the players for their actions in Γ. Punishment is not restricted to players of Γ, because morality is defined over *actions* of players, irrespective of whom the action was directed at, and because of the complete transparency of individual behaviour within a very small group. The actions that are punished by player i are

[5] Although cognitive capacities *did* increase in the course of human evolution, this process is assumed to be too slow to interact with the dynamics in this model.

[6] Restrictions $0 < a$ and $0 < b < 1$ ensure that Γ is indeed a prisoners dilemma. Parameter a is strictly smaller than 1 so that unilateral defection of one payer does not generate higher total payoff than bilateral cooperation. If there is only one player in the group, he receives a payoff of b, with $0 < b < 1$ and the game Γ^e is over.

[7] The assumption of observation with certainty is inessential. The main results of the model can be proven for the case that players not playing Γ themselves do not observe the actions in Γ for sure but only with a sufficiently high probability $0 < \gamma < 1$.

	C	**D**
C	$1,\ 1$	$0,\ 1+a$
D	$1+a,\ 0$	$b,\ b$

$$0 < a, b < 1$$

Figure 4.1: The prisoner's dilemma game Γ

called player i's *moral code*[8]

$$M_i \in \{\{\emptyset\}, \{C\}, \{D\}, \{C, D\}\}$$

If the action of player j playing Γ falls into the moral code of player i, then player i punishes player j by imposing a cost α on the perpetrator. The individual cost that player i has to bear for punishing action s of player j is denoted by $\beta(s_j)$ and given by equation 4.2. While the moral code of player i determines *what* he punishes, the individual *cost* of punishing strategy s_j is determined by $n^P(s_j)$, the total number of agents who punish s_j. If the number of players punishing the same action at the same time is above a critical level n_c^P, the individual cost of punishment falls to some positive but very small ε. I assume that $\alpha > \beta_0$ and that $n_c^P \leq n_{max}$.

$$\beta(s_j) = \begin{cases} \beta_0 & \text{with } \varepsilon < \beta_0 < \alpha \quad \text{if } n^P(s_j) < n_c^P \\ \varepsilon & \text{with } \varepsilon < \frac{\alpha}{n_{max}} \quad \text{otherwise} \end{cases} \tag{4.2}$$

with

$$n^P(s_j) = \sum_{k=1}^{n_t} \mathbf{1}_{s_j \in M_k} \tag{4.3}$$

denoting the number of agents that will punish action s in Γ.

[8] Kuzmics & Rodriguez-Sickert (2007) also use the concept of "moral codes". In contrast to Kuzmics & Rodriguez-Sickert's model, however, player i does not need to be an active participant in Γ to observe and potentially punish actions of player j.

For simplicity, I assume that the punishment technology, including all four parameters α, β_0, ε and n_c^P, is constant over all agents and actions in Γ.

A strategy s^e in Γ^e consist of a choice between cooperation and defection in Γ and a moral code M to be acted upon in the punishment stage. Let the set of all those strategies be denoted by S^e.

Payoffs in Γ^e: For each player i selected to play Γ^e against another player j and each strategy $s^e \equiv \{s, M\} \in S^e$ payoffs are given by:

$$y_i^e(s_i^e, s_j^e) = y_i(s_i, s_j) - \alpha\, n_{\setminus i}^P(s_i) - \beta(s_j)\, \mathbf{1}_{s_j \in M_i} \tag{4.4}$$

with

$$n_{\setminus i}^P(s_i) = \sum_{k=1, k\neq i}^{n_t} \mathbf{1}_{s_i \in M_k} \tag{4.5}$$

Equation 4.4 shows that under actual punishment equilibrium payoffs y in Γ are reduced for an agent both by punishment received and by the cost of (potentially) punishing the other player.

For players k *not* playing Γ, payoffs are:

$$y_k^e(s_i^e, s_j^e) = -\left(\beta(s_i)\, \mathbf{1}_{s_i \in M_k} + \beta(s_j)\, \mathbf{1}_{s_j \in M_k}\right) \tag{4.6}$$

Players not participating in Γ still bear the cost of punishing if the actions of players in Γ are in their moral code.

Markov perfect, pure strategies: Each agent selects a *pure* strategy $s_i^e \in S^e$ to maximize his payoff y_i^e every time Γ^e is played. He can only condition his choice of strategy on the current state of the group in a given period. The state consists of group size n_t and the number of players with one of the 4 types described below.

Types of genetically determined behaviour: There are four possible types of players in the group. Let the type of player i be denoted by $\sigma_i \in \{N, PC, PD, PA\}$. The number of players of each type in period t are indicated by $n_t^\sigma = n_t^N, n_t^{PC}, n_t^{PD}$ and n_t^{PA}, respectively. I use the same superscripts for other variables that differ by type, too.

All four types are free to chose their actions s in Γ. Their moral codes, however, are (partly) genetically determined. 3 of the 4 types are getting angry when observing defection or/and cooperation and have no choice but to punish the agent playing this

action.[9] If the behaviour of an agent in the punishment stage is not completely genetically determined, he is free to include the remaining action(s) into his moral code or not.

N-type Non-determined. Players of type N are not at all genetically determined. They freely select their actions in both stages of Γ^e.

PC-type Punisher of cooperation. Free to select $s \in S$ in Γ, but gets upset and for sure punishes every player who cooperates in Γ.

PD-type Punisher of defection. Free to select $s \in S$ in Γ, but gets upset and for sure punishes every player who defects in Γ.

PA-type Punisher of all actions. Free to select $s \in S$ in Γ, but uses every chance to harm other players in the punishment stage, irrespective of their actions in Γ.

I assume that types are common knowledge for all players in the group.[10] The initial situation is assumed to be one with *no* genetic determination. Initially, mutations happen in a world populated by N-types.

Noise in behaviour: With probability $0 < \eta < \frac{1}{2}$, agent i plays some action $s' \in S$ although he had chosen to play $s \neq s'$ in Γ. He wants to cooperate, but defects or vice versa. Errors in Γ are equally likely for all types of players. I assume that there is no noise in the punishment stage and that the amount of noise is strictly positive but small.[11]

Reproduction, selection and mutation: At the end of every period reproduction happens. Reproductive success is measured as the number of offspring for each individual. It is based on individual fitness - the average payoff of each player in period t. Selection happens in the following way: The most successful individual(s), i.e., those group members with maximum average payoff in t, reproduce twice, at the expense of the same number of players with lowest average payoffs, who do not reproduce at all. All other players reproduce once. In case more than half the members

[9] Alternatively, one can express this assumption as a utility function that includes an infinite psychological cost due to observing an action s that is in the moral code of the agent. This cost can only be compensated by punishing the offender in the extent α.

[10] This assumption is an innocuous simplification within the small group, continuous interaction model. It will be clear later on, that behaviour in the punishment stage is completely revealing. Only players who are genetically determined to punish will do so. As Γ^e is played infinitely often in every period and the number of players is finite, types could be learnt arbitrarily quickly.

[11] The maximum level of noise will be specified relative to other model parameters later on. For sure $\eta < \frac{1}{2}$, so that a player is always more likely to play the action he selected than the one he did not.

of the group share the same, maximum payoff, the players who reproduce twice are randomly selected, and only players with maximum payoff reproduce at all. In all cases, group size n_t does not change due to reproduction and selection dynamics.

Every player passes on his type to his offspring with probability $(1-\mu)$, which, by assumption, is very close to 1. With a small probability μ, however, a mutation happens and the newborn changes his type. For simplicity I assume that the transition probability is the same for any pair of types.

Population size dynamics: Two processes govern the development of group size n_t - intergroup contact and changes to the environment. It is assumed that the number of groups that have a chance to interact with each other is *finite*. Let this set of groups be called the *population*. For simplicity it is assumed that the entire population lives in a common habitat, ie., that the environmental conditions for all groups in a population are the same. All parts of the common habitat are affected by environmental shocks in the same way.

With frequency $f_{int} > 0$, intergroup contact happens within a given habitat. How group size changes in the event of intergroup contact depends on strategies played in Γ^e. If two groups confront each other, the group with lower average equilibrium payoff gets eradicated. The other group will grow by assimilating all foreign females. For simplicity, I assume that the winning group replicates itself - a second group is created with exactly the same size and composition of types than the original winning group.[12]

If the individuals in both groups receive the same average payoff, no change in group size or composition happens for either group.

Formally, let two groups confronting each other be labelled group A and B. Group size dynamics due to intergroup contact within a given habitat is assumed to depend on average payoff in period t over all players in group A ($\bar{\bar{y}}_t^A$) and group B ($\bar{\bar{y}}_t^B$).

The new size of group A in period $t+1$ is given by equations 4.7 and 4.8. Group size dynamics for group B are analogical.

$$n_{t+1}^A = \begin{cases} 0 & \text{if} \quad \bar{\bar{y}}_t^A < \bar{\bar{y}}_t^B \\ n_t^A & otherwise \end{cases} \tag{4.7}$$

Let m indicate the number of groups in a given geographical area, and $m_t^A = m_t^B$ be the number of groups at time t with exactly the characteristics of group A and B,

[12] This assumtion is harmless for an environment where the original groups are of similar size; and this group size is already exhausting the habitat's capacity or the players ability to stay together as a group while feeding and traveling, i.e., $n_t \approx \min\{N_{max}, n_{max}\}$ for all groups.

respectively. Then

$$m_{t+1}^A = \begin{cases} m_t^A - 1 & \text{if} \quad \bar{y}_t^A < \bar{y}_t^B \\ m_t^A + 1 & \text{if} \quad \bar{y}_t^A > \bar{y}_t^B \\ m_t^A & otherwise \end{cases} \tag{4.8}$$

Changes to the quality of the environment occur with frequency f_{env}. These shocks are completely exogenous to the behaviour of group members. A shock at time t determines individual group size and maximum carrying capacity of the group's habitat N_{max}. Maximum group size n_{max} does not depend on the environment and remains unchanged.

For simplicity only two, alternating, states of nature are considered. In the bad environment habitats have a carrying capacity $N_{max} = N^L$. In the good environment it is higher with $N_{max} = N^H$. Equations 4.9 and 4.10 specify the transition dynamics between the two states.

$$\left.\begin{array}{rcl} N_{max}(t+1) & = & N^H \geq n_{max} \\ n_{t+1} & = & n_{max} \end{array}\right\} \begin{array}{l} \text{transitioning} \\ \text{from } N^L \text{ to } N^H \end{array} \tag{4.9}$$

$$\left.\begin{array}{rcl} N_{max}(t+1) & = & N^L < n_{max} \\ n_{t+1} & = & \min\{n_t, N^L\} \end{array}\right\} \begin{array}{l} \text{transitioning} \\ \text{from } N^H \text{ to } N^L \end{array} \tag{4.10}$$

Changes in group size are referring to all groups in a given environment.

If the group's habitat changes from bad to good quality and there are different types of players in the group, the additional players in period $t+1$ are of the same type as the fittest individual(s) in period t, immediately before the positive shock to the environment. If group size exceeds the habitat's new maximum carrying capacity in the event of a bad shock to the environment, $N^L - n_t$ of the least successful individuals die.

4.3 Static analysis

This section analyzes optimal choices and relative payoffs for the different types of players in *one* period. Group size and composition with respect to type are constant. The time index t is dropped to simplify notation.

4.3.1 Optimal choice

The first lemma establishes that in the model presented in this chapter there is no punishment without genetic determination.

Lemma 1 (No voluntary punishment). *No player will ever choose to add an action s to his moral code.*

Proof. This lemma is a direct consequence of the fact that holding a non-empty moral code is a weakly dominated strategy.

Choice of action in Γ is independent of previous behaviour of other players within the same period because players can only condition their choice of strategy on group size and composition with regards to type. Both of those state variables are constant within a given period. Therefore, any trigger strategies are prevented that threaten to punish undesired behaviour in Γ by choosing a different action s the next time Γ is played. In addition, no player can credibly commit on punishing any action s in the second stage of Γ^e unless he is genetically determined to do so. Therefore, no player can deter any other player from playing an action s that is not in his moral code. Formally, for every strategy $s^e \equiv (s, M)$ with $M \neq \emptyset$ there is an alternative strategy $s^{e\prime} \equiv (s, \emptyset)$ that implies the same action in Γ in combination with an empty moral code. Holding a non-empty moral code M leads to actual punishment if an action $s \in M$ is played in Γ. But this just decreases individual payoffs without any benefit. Therefore s^e is weakly dominated by $s^{e\prime}$.

Moreover, for any positive amount of noise ($\eta > 0$), holding a non-empty moral code is strictly dominated, because both cooperation and defection are played in equilibrium and costly punishment happens for any $M \neq \emptyset$. ■

An immediate consequence of lemma 1 is that behaviour in the punishment stage is determined by player i's own type only. Player i will punish action s in Γ if and only if nature determined him to do so.

The next lemma gives a sufficient condition for cooperation in Γ in terms of the distribution of types in the group.

Lemma 2 (Sufficiency for cooperation in Γ). *Let*

$$n^{\sigma}_{\setminus i} = n^{\sigma} - \mathbf{1}_{\sigma_i = \sigma}$$

denote the number of players of type σ, excluding player i himself. Then player i will chose cooperation in Γ if

$$n^{PD}_{\setminus i} - n^{PC}_{\setminus i} \geq \frac{\max\{a, b\}}{\alpha} \tag{4.11}$$

i.e., the number of others in the group with type PD, who are genetically determined to punish defection, but not cooperation, exceeds the number of other players with type PC, who are determined to punish only cooperation, by an amount that is proportional to the maximum advantage of defection in Γ and inversely proportional to the strength of punishment α.

Proof. Let $\Delta_i^e(C,D)$ denote the difference in payoff for player i between cooperation and defection in Γ. Player i choses cooperation when playing against player j (taking his strategy as given) iff

$$\Delta_i^e(C,D) \equiv y_i^e(C, s_j^e) - y_i^e(D, s_j^e) \geq 0$$

But

$$y_i^e(C, s_j^e) = y_i(C, s_j) - \alpha\left(n_{\setminus i}^{PC} + n_{\setminus i}^{PA}\right) - \beta \mathbf{1}_{s_j \in M_i}\,, \tag{4.12}$$

$$y_i^e(D, s_j^e) = y_i(D, s_j) - \alpha\left(n_{\setminus i}^{PD} + n_{\setminus i}^{PA}\right) - \beta \mathbf{1}_{s_j \in M_i} \tag{4.13}$$

and

$$y_i(C, s_j) - y_i(D, s_j) = \begin{cases} a & \text{if } s_j = C \\ b & \text{if } s_j = D \end{cases} \tag{4.14}$$

Subtracting equation 4.13 from equation 4.12 and substituting equation 4.14 results in

$$\Delta_i^e(C,D) = \alpha\left(n_{\setminus i}^{PD} - n_{\setminus i}^{PC}\right) \geq \begin{cases} a & \text{if } s_j = C \\ b & \text{if } s_j = D \end{cases} \tag{4.15}$$

If (4.15) holds for both cooperation and defection by player j, cooperation is ensured. In this case, the last line can easily be rearranged to obtain inequality 4.11. ■

Corollary 1. *Player i will select defection in Γ if*

$$n_{\setminus i}^{PD} - n_{\setminus i}^{PC} < \frac{\min\{a,b\}}{\alpha} \tag{4.16}$$

i.e., defection remains the dominant strategy if the extent of punishment is insufficient to counterbalance the advantage of defection under both strategies of the opponent j.

4.3.2 Relative fitness

Proposition 1 presents the first main result of this chapter. Let $n_{\setminus i}^{\sigma}$ again denote the number of players of type σ in the group without player i.

Proposition 1 (Relative fitness in small groups). *In a group of players who all choose the same action s in Γ but make mistakes with strictly positive probability, i.e.,*

$$\begin{aligned} n_{\setminus i}^{PD} - n_{\setminus i}^{PC} &\geq \frac{\max\{a,b\}}{\alpha} \quad \forall i \in \{1,..,n\} \\ &\textbf{or} \\ n_{\setminus i}^{PD} - n_{\setminus i}^{PC} &< \frac{\min\{a,b\}}{\alpha} \quad \forall i \in \{1,..,n\} \end{aligned} \tag{4.17}$$

and

$$0 < \eta < \frac{1}{2}, \tag{4.18}$$

genetically determined punishers have an advantage if the group is sufficiently small. A group is sufficiently small if total *group size is restricted by*

$$n < \frac{\alpha}{\beta} + 1 \tag{4.19}$$

Proof. Let

$$\mathbf{1}_{\sigma_j} = \begin{cases} 1 & \text{if } \sigma_i = \sigma_j \\ 0 & \text{otherwise} \end{cases}$$

be an indicator function of player i's type.

If for all players $n_{\setminus i}^{PD} - n_{\setminus i}^{PC} < \frac{\min\{a,b\}}{\alpha}$, every player selects defection, and the average payoff for player i is given by equation 4.20.

$$\bar{y}_i^e = \frac{2}{n}\,\bar{y}_{i,1}^e + \frac{n-2}{n}\,\bar{y}_{i,2}^e \tag{4.20}$$

Equation 4.20 shows that payoff for player i is the average of two situations. Player i is selected to play Γ with frequency $f(i \text{ plays } \Gamma) = \frac{1}{n} + \frac{1}{n-1}\frac{n-1}{n} = \frac{2}{n}$.[13] In this case i receives payoff $\bar{y}_{i,1}^e$. If i is just an observer, which happens with frequency $\frac{n-2}{n}$, he receives $\bar{y}_{i,2}^e$. Taking into account equations 4.4 and 4.6, and a positive amount of noise in behaviour,

[13] The first term represents the frequency that i is selected as first player in Γ and the second that he is selected as the second player, conditional on not already being selected as the first.

average payoffs can be written as

$$
\begin{aligned}
\bar{y}_i^e \;=\; & \frac{2}{n}\Bigg\{ (1-\eta)^2\Big[y(D,D) - \alpha(n_{\setminus i}^{PD} + n_{\setminus i}^{PA}) - \beta(\mathbf{1}_{PD} + \mathbf{1}_{PA})\Big] \\
& \quad + \eta\,(1-\eta)\,\Big[y(C,D) - \alpha(n_{\setminus i}^{PC} + n_{\setminus i}^{PA}) - \beta(\mathbf{1}_{PD} + \mathbf{1}_{PA})\Big] \\
& \quad + (1-\eta)\,\eta\,\Big[y(D,C) - \alpha(n_{\setminus i}^{PD} + n_{\setminus i}^{PA}) - \beta(\mathbf{1}_{PC} + \mathbf{1}_{PA})\Big] \\
& \quad + \eta^2\,\Big[y(C,C) - \alpha(n_{\setminus i}^{PC} + n_{\setminus i}^{PA}) - \beta(\mathbf{1}_{PC} + \mathbf{1}_{PA})\Big]\Bigg\} \\
& - \frac{n-2}{n}\,\beta\,\Big\{2(1-\eta)^2(\mathbf{1}_{PD} + \mathbf{1}_{PA}) \\
& \qquad\qquad + \eta\,(1-\eta)(\mathbf{1}_{PC} + \mathbf{1}_{PA} + \mathbf{1}_{PD} + \mathbf{1}_{PA}) \\
& \qquad\qquad + (1-\eta)\,\eta(\mathbf{1}_{PD} + \mathbf{1}_{PA} + \mathbf{1}_{PC} + \mathbf{1}_{PA}) \\
& \qquad\qquad + 2\,\eta^2\,(\mathbf{1}_{PC} + \mathbf{1}_{PA})\Big\}
\end{aligned}
\tag{4.21}
$$

The first four lines in equation 4.21 sum up the expected payoff to player i when he is selected to play Γ. Each of the four lines reflects another pair of actions that are actually played due to noise in behaviour. $y(s_i, s_j)$ denotes payoffs in Γ. Player i's strategy is punished by all players who have i's strategy in their moral codes, except of i himself. An additional cost of punishing occurs when player i has s_j, the strategy of this opponent in Γ, in his moral code.

If player i does not play Γ, his cost is determined by his own type and the actions actually being taken in Γ. Lines 5 to 8 in equation 4.21 sum up the average cost of punishing when i is only an observer.

Simplifying leads to:

$$
\begin{aligned}
\bar{y}_i^e &= \frac{2}{n}\Bigg\{\underbrace{(1-\eta)^2 y(D,D)+\eta(1-\eta)y(C,D)+(1-\eta)\eta y(D,C)+\eta^2 y(C,C)}_{\bar{y}_D^\Gamma} \\
&\quad -\alpha\Bigg[n_{\setminus i}^{PA}\Big(\underbrace{(1-\eta)^2+2\eta(1-\eta)+\eta^2}_{1}\Big)+n_{\setminus i}^{PD}\Big(\underbrace{(1-\eta)^2+(1-\eta)\eta}_{1-\eta}\Big)+n_{\setminus i}^{PC}\Big(\underbrace{\eta(1-\eta)+\eta^2}_{\eta}\Big)\Bigg] \\
&\quad -\beta\Bigg[\mathbf{1}_{PA}\Big(\underbrace{(1-\eta)^2+2\eta(1-\eta)+\eta^2}_{1}\Big)+\mathbf{1}_{PD}\Big(\underbrace{(1-\eta)^2+\eta(1-\eta)}_{1-\eta}\Big)+\mathbf{1}_{PC}\Big(\underbrace{(1-\eta)\eta+\eta^2}_{\eta}\Big)\Bigg]\Bigg\} \\
&\quad -\frac{n-2}{n}\beta\Bigg\{\mathbf{1}_{PA}\Big(\underbrace{2(1-\eta)^2+4\eta(1-\eta)+2\eta^2}_{2}\Big)+\mathbf{1}_{PD}\Big(\underbrace{2(1-\eta)^2+\eta(1-\eta)+(1-\eta)\eta}_{2(1-\eta)}\Big) \\
&\qquad +\mathbf{1}_{PC}\Big(\underbrace{\eta(1-\eta)+(1-\eta)\eta+2\eta^2}_{2\eta}\Big)\Bigg\} \\
&= \frac{2}{n}\bar{y}^\Gamma-\frac{2\alpha}{n}\Big[n_{\setminus i}^{PA}+(1-\eta)n_{\setminus i}^{PD}+\eta n_{\setminus i}^{PC}\Big]-\Big[\frac{2\beta}{n}+\frac{2(n-2)}{n}\beta\Big]\Big(\mathbf{1}_{PA}+(1-\eta)\mathbf{1}_{PD}+\eta\mathbf{1}_{PC}\Big) \\
&= \underbrace{\frac{2}{n}\Big[\bar{y}^\Gamma-\alpha\Big[n^{PA}+(1-\eta)n^{PD}+\eta n^{PC}\Big]\Big]}_{\bar{y}_i^{e,N}}+\frac{2}{n}\Big[\Big[\alpha-(n-1)\beta\Big]\Big(\mathbf{1}_{PA}+(1-\eta)\mathbf{1}_{PD}+\eta\mathbf{1}_{PC}\Big)\Big] \\
\bar{y}_i^e &= \bar{y}_i^{e,N}+\frac{2}{n}\Big(\alpha-(n-1)\beta\Big)\Big[\mathbf{1}_{PA}+(1-\eta)\mathbf{1}_{PD}+\eta\mathbf{1}_{PC}\Big] \qquad (4.22)
\end{aligned}
$$

To get to the penultimate line above, I use the fact that $n_{\setminus i}^{\sigma}=n^{\sigma}-\mathbf{1}_{\sigma}$. The average payoff $\bar{y}_i^{e,N}$ for a player with type N in equation 4.22 follows from the penultimate line with $\mathbf{1}_{PA}=\mathbf{1}_{PD}=\mathbf{1}_{PC}=0$.

If all players select cooperation in Γ then average payoffs are given in the same way, only that noise in behaviour now works in the opposite direction:

$$
\begin{aligned}
\bar{y}_i^e &= \bar{y}_i^{e,N}+\frac{2}{n}\Big(\alpha-(n-1)\beta\Big)\Big[\mathbf{1}_{PA}+\eta\mathbf{1}_{PD}+(1-\eta)\mathbf{1}_{PC}\Big] \qquad (4.23)\\
&\text{with}\\
\bar{y}_i^{e,N} &= \frac{2}{n}\Big[\bar{y}_C^\Gamma-\alpha\Big[n^{PA}+\eta n^{PD}+(1-\eta)n^{PC}\Big]\Big]\\
&\text{and}\\
\bar{y}_C^\Gamma &= (1-\eta)^2 y(C,C)+\eta(1-\eta)y(D,C)+(1-\eta)\eta\, y(C,D)+\eta^2 y(D,D)
\end{aligned}
$$

In both equation 4.22 and 4.23 the sum in square brackets is positive for any player i who is genetically determined to punish ($\sigma_i\in\{PA,PD,PC\}$), because of the level of noise is strictly positive, but small ($0<\eta<\frac{1}{2}$).

A genetically determined punisher has a payoff advantage over the N-type if the term in big round brackets is positive, which is the case iff

$$\Big(\alpha - (n-1)\,\beta\Big) > 0 \quad \Leftrightarrow \quad \alpha > (n-1)\,\beta \quad \Leftrightarrow \quad n < \frac{\alpha}{\beta} + 1 \tag{4.24}$$

■

The intuition behind proposition 1 is very simple and can be seen best in the second term in equation 4.24. If all players receive the same payoff in Γ, a genetically determined punisher has a *relative* advantage if the benefit (α) of being punished by one player less (himself) overcompensates the cost of punishing the $n-1$ other players in the group.

Note that proposition 1 only depends on *total* group size and is independent of relative frequency of types within the group.

Looking at the sums of indicator variables in square brackets in equations 4.22 and 4.23, one can immediately recognize a ranking in relative fitness between the three genetically determined types.

Corollary 2 (Ranking between punishers). *Assume that group size is such that genetically determined punishers have a payoff advantage. Then individuals who punish all the time do best for any equilibrium strategy in* Γ*. Furthermore, if noise is small* ($0 < \eta < \frac{1}{2}$)*, then punishers of defection do better (worse) than punishers of cooperation if all players choose defection (cooperation).*

$$\bar{y}_i^{e,PA} > \bar{y}_i^{e,PD} > \bar{y}_i^{e,PC} > \bar{y}_i^{e,N} \quad \textit{if} \quad n_{\setminus i}^{PD} - n_{\setminus i}^{PC} < \frac{\min\{a,b\}}{\alpha} \quad \forall i, \tag{4.25}$$

$$\bar{y}_i^{e,PA} > \bar{y}_i^{e,PC} > \bar{y}_i^{e,PD} > \bar{y}_i^{e,N} \quad \textit{if} \quad n_{\setminus i}^{PD} - n_{\setminus i}^{PC} \geq \frac{\max\{a,b\}}{\alpha} \quad \forall i \tag{4.26}$$

Corollary 2 simply says that, when punishing is advantageous, those who punish most frequently derive the highest benefits.

4.4 The evolution of moral punishment

The main section of this chapter analyzes the dynamics in the evolution of moral punishment. It shows that, under conditions that were typical for our human ancestors, evolution necessarily resulted in the prevalence of PD-types and cooperation as the unique norm shared by all players. More precisely:

Proposition 2 (The dominance of cooperation). *Consider an environment that changes in quality between the good and the bad state, ie.* $f_{env} > 0$.

In such a volatile environment, all players in a population will eventually be punishers of defection. Mutants of other types do not reproduce. Cooperation will be the norm shared by all non-mutant players.

The only temporary exception can be PA-types in an isolated group.

Before subsections 4.4.1, 4.4.2 and 4.4.3 provide formal prove, the next paragraphs will give an overview over the evolutionary dynamics in verbal form.

According to proposition 2, selection dynamics will ensure that any sequence of events[14] leads to a population of only PD-types. The result can be obtained if and only if the environment changes in quality, alternating between the bad and the good state. Moreover, a *population* of only punishers of defection is evolutionary stable against single mutations.

What are the principal[15] dynamics leading to this result?

By assumption, evolutionary dynamics are starting in an original population of only N-types. The first mutant, who is genetically determined to punish any action s in Γ, will survive if the group is sufficiently small - depending on the punishment technology as expressed in proposition 1.

Small enough groups will occur only in the bad state of nature. To simplify the argument, it is assumed that all groups in the bad state will be small enough for genetically determined punishers to have a relative advantage.[16]

Therefore, a mutant of type PC, PD or PA will reproduce more successfully than the N-types in any of the small groups under the bad state of nature. This is illustrated in figure 4.2, points A and C.

[14] Except of very rare events like simultaneous mutations of multiple individuals, which are not considered in this model.

[15] Only the main line of argument is given here - for a detailed account of all special cases and specific assumptions the reader is referred to subsections 4.4.1, 4.4.2 and 4.4.3.

[16] This applies to any mutants in groups of only N-types. For other type combinations, the advantage lies with the strongest punisher according to corollary 2, equation 4.25.

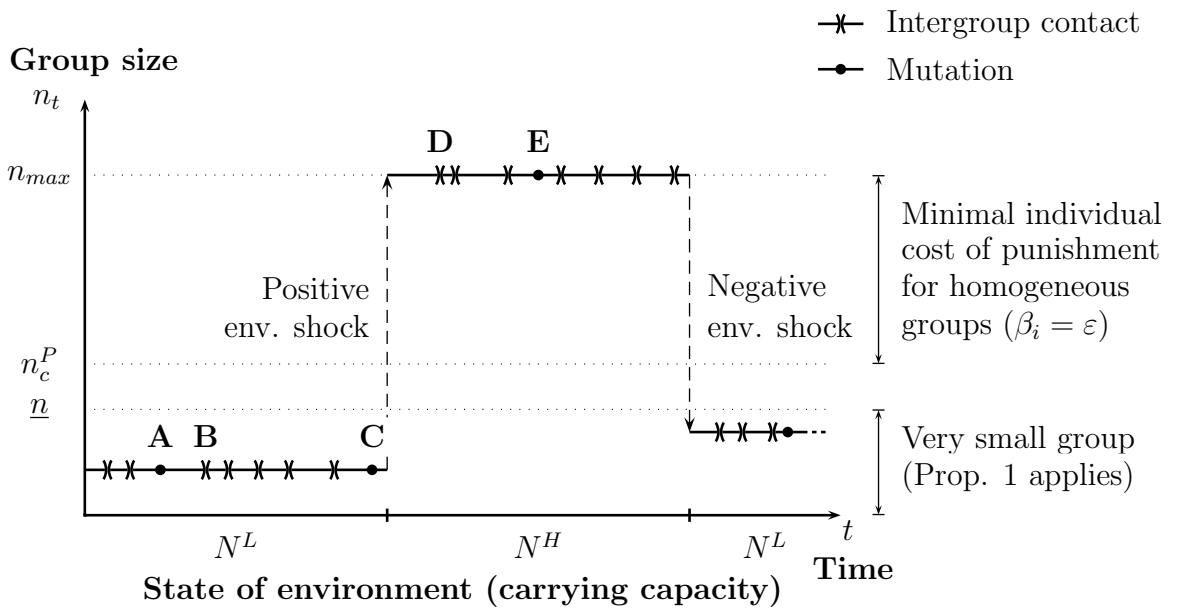


Figure 4.2: Evolution of moral punishment - illustrative timeline

As long as the group does not change in size due to a shock to the environment, the mutants retain their relative advantage. Because groups are of finite size, the replacement process finishes in finite time and mutants fix their types - if the small group remains isolated. The term *fixation* means that all N-types have been replaced by the mutant type, so that all players are genetically determined to punish the same action s in Γ^e.

However, genetically determined punishers only have a *relative* advantage over the N-types *in their group*. In *absolute* terms, punishment reduces the payoffs for all members of the group, given that the same strategy is played in Γ then without punishment. But in a stable bad environment, defection is the strategy of choice in all groups.[17] Therefore, all groups that contain mutants will have lower average payoffs than homogeneous groups of non-punishers in the bad environment. But, by assumption, intergroup contact will eradicate any small group with inferior average payoff. Contact between N-type and mutant groups happens eventually and mutants will die out. This will be the case for the group illustrated in figure 4.2, that contains a number of mutants after initial mutation at A, but encounters another group at point B.

[17] In general, it is assumed that the number of punishers of defection in very small groups is not sufficient for cooperation to be enforced. This loads the dice against the emergence of cooperation as the dominant norm.

But what happens if the environment improves to the good state while there is at least one group left that contains mutants, for instance the one that was created at point C in figure 4.2?

If the environment changes from bad to good state, groups become big. By assumption they reach n_{max}, the maximum possible group size. When a group increases in size, the additional individuals are of the same type as the fittest member(s) of the original small group - the genetically determined punishers. Thus, when the mutants were close to fixing their type in the small group, they constitute the overwhelming majority in the big group. It is assumed that, in general, their number in the big group will exceed n_c^P and the cost of punishment $\beta(s_i)$ will reduce to ε because a sufficient number of genetically determined players punish the same actions at the same time.[18] This way, mutants retain their *relative* advantage, continue to replace the remaining N-types, and finally fix their type. If a group was homogeneous in type already in the bad state, it will also be homogeneous after the change to the good environment, of course.

The next important fact to notice is that homogeneous large groups are evolutionary stable against single mutants of any type - *irrespective of the type* of the existing group members. Why is this? The reason is again the low cost of punishment when behaviour is perfectly coordinated between a large set I of genetically determined individuals. Every member $i \in I$ of this set has a *relative* advantage against individuals of all other types in his group. Proposition 1 applies with $\beta_i = \varepsilon$. A punisher $j \notin I$ faces high cost of punishment as he does not punish the same action as the other punishers in the group. For example, a mutant j of type PA incurs cost $\beta_j = \beta_0$ in a group of PD or N-types, because he is the only one punishing cooperation. But single mutants with a cost of punishment of β_0 have a relative advantage in very small groups only. In a group of size n_{max}, the mutant will not reproduce. To refer back to figure 4.2, a mutation at point E in a large group will have no lasting effect.

But what about intergroup contact? If a group containing any number of mutants encounters a group of only N-types in the bad state of nature, for instance at point B in figure 4.2, the mutants are eradicated. In large groups, the same applies to groups of PC and PA-types. In these groups, defection is still the common strategy of choice and the additional cost of punishment reduces average payoffs below the level within pure N-type groups.

[18] In case the number of individuals are insufficient for this 'social cost effect' of punishment to apply, the punishers will not survive in the big group.

This is not the case for groups of PD-types, who punish defection. In these groups, players can be deterred from choosing defection and cooperation is enforced.[19] Average payoff in Γ^e increases over payoff in groups of only N-types, in which no punishment happens, but defection is the strategy of choice. This is the case for low levels of noise, so that the payoff-advantage of the cooperation equilibrium is not over-compensated by the cost of punishing accidental defection of PD-type players.

Therefore, groups consisting only of punishers of defection prevail whenever they encounter any other group in the good environment, for instance at point D in figure 4.2. Over time, groups of N, PC and PA types are replaced by PD-type groups, whenever they encounter a group of punishers of defection. If the good state of the environment lasts long enough, punishers of defection fix their type in the entire population.

The last step to conclude the dominance of cooperation is to understand, why a *population* of PD-types is evolutionary stable under changing environmental conditions. In the good environment, cooperation is the strategy of choice for all players, including mutants. Punishers of defection have a *relative* advantage against other types due to the low cost of coordinated punishment - proposition 1 applies with $\beta = \varepsilon$.

In the bad environment, defection is chosen and PD types are stronger punishers than PC and N-types (corollary 2). Consequently, they eliminate single mutants of those two types. Only a mutant of type PA will survive and fix his type in a group of only PD-types. However, a group of PA-types will not survive intergroup contact. PA-type groups will always have a lower average payoff than groups of PD-types, because they punish both cooperation and defection[20] and, thus, incur higher cost of punishment. Therefore, only PD-types will survive and dominate the population in the long run, and cooperation is the unique norm shared by all individuals.

The following subsections provide the formal proof of proposition 2. The argument proceeds in several steps, organized as individual lemmata. Every time, a mutation is considered that happens in a group of players with a given distribution of types. A single mutant is introduced, who differs in type, i.e., who carries a moral code that is unique in the group. To simplify notation the subscript i for an individual player is dropped from now on.

[19] This applies for sufficiently high levels of punishment α and for actions that are played as chosen, ie., except of defection mistakenly played due to noise in behaviour $\eta > 0$.

[20] For the same reason they cannot enforce cooperation in the good state.

First a baseline scenario is considered, in which a single mutant is introduced into a group of only N-types. The first subsection analyzes the evolutionary dynamics under constant group size and no additional mutations. The analysis shows that mutants of any type are successful if the group is small enough. Furthermore, under conditions specified below, the entire group will be of the mutant's type within finite time, i.e., there will be a fixation of types. Finally, the result for the baseline scenario can be adapted to groups of any initial composition with regard to types.

The second subsection discusses the question of evolutionary stability, in particular in big groups. It will be shown that a group that is homogeneous in type is robust against mutations as long as it is *big* enough. Small groups are only robust against mutations if they consist of PA-types, medium sized ones if they consist of N-types. Subsections 4.4.1 and 4.4.2 will make precise what is meant by 'small', 'medium' and 'big'. Finally, subsection 4.4.3 will look at the population dynamics and clarify the conditions under which punishment of defection emerges as the only outcome stable against mutations, intergroup contact and environmental shocks.

4.4.1 Initial mutation and fixation of types

First the analysis turns to a mutation that happens in a group of N-types, i.e., players who are not genetically determined to punish any strategy s in Γ. Let this event be called the 'initial mutation' and the period it happens in the 'initial period', denoted t_0.

Lemma 3 (Success of initial mutant). *Consider a group of N-types. A mutant who is genetically determined to punish will reproduce twice in this group after the initial period if*

1. *Group size n at the time of initial mutation t_0 is restricted by*

$$n_{t_0} < \frac{\alpha}{\beta_0} + 1 \tag{4.27}$$

2. *And either of the following two conditions on the punishment technology, in particular on the extent of punishment α, are true:*

   (a) *The mutant can not threaten to punish defection (enough) to induce any change in behaviour* **or**

   (b) *the punishment technology allows the mutant to punish defection so strongly that cooperation turns into a dominant strategy for the N-types.*

Proof. According to lemma 1, all players in a homogeneous group of N-types will play defection. Because there is only one mutant in the group, the scale effect in the punishment technology does not apply and the cost of punishment is given by $\beta = \beta_0$. If a mutant enters the group, choice of strategy in Γ might change, but only if the mutant threatens to punish defection sufficiently strongly.

Condition 2a: If the mutant is of type PC or PA, corollary 1 holds and all group members will continue playing defection in Γ. The same is true if the mutant is of type PD, but the amount of punishment α is sufficiently low. If $\alpha < \min\{a, b\}$ corollary 1 holds as well, for both N-types and, of course, the mutant. Again, all players will continue playing defection.

Thus, if condition 2a holds, all players in the group will play the same strategy in Γ (defection) and proposition 1 holds. Therefore, the mutant has an advantage and reproduces twice if the group is sufficiently small, i.e., $n_{t_0} < \frac{\alpha}{\beta_0} + 1$

Condition 2b: If, however, the mutant is of type PD and punishment is sufficiently strong, i.e., $\alpha \geq \max\{a, b\}$, proposition 1 does not apply. This is because all players but the mutant are facing the threat of punishment and might choose to cooperate, while the mutant does not punish himself and, therefore, continues to play defection.

If and when the N-types will cooperate depends on the size of α relative to the parameters a and b. The payoffs in Γ^e for the different pairs of player types are shown in figures 4.3 and 4.4. Equilibrium payoffs are shown in bold face.

Case 1:
$\alpha \geq \max\{\mathbf{a}, \mathbf{b}\}$

		C	**D**
	C	1, 1	**0, >1**
	D	≤ 1, 0	≤0, >0

N vs. PD

	C	**D**
C	**1, 1**	0, ≤ 1
D	≤ 1, 0	≤ 0, ≤ 0

N vs. N

Figure 4.3: Payoff in Γ^e if one mutant of type PD enters a group of N-types - cooperation becomes the dominant strategy for N-types.

Figure 4.3 shows the case of very strong punishment of defection. A mutant of type PD enters the group of N-types and punishes any player for defection by an amount that is larger than the benefit of defecting, irrespective of the strategy chosen by the second player in Γ. Thus, cooperation becomes the dominant strategy for the N-types. Condition 2b is fulfilled.

If, in this situation, the mutant of type PD chose cooperation as well, he would have an advantage because condition 1 of lemma 3 ensures sufficiently small groups for propo-

sition 1 to apply. However, the mutant does not punish himself and, thus, has no reason to cooperate. To the contrary, the PD-type increases payoffs by defecting, over what he would have achieved with cooperation.

Let $\bar{y}_C^{e,PD}$ be the average payoff of the mutant PD-type in period t_0 if he cooperates in the same way as all the N-types. Use $\bar{y}_D^{e,PD}$ for the mutant's payoff if he defects in this situation. Let $\Delta_{D,C}$ be the difference between the two. Equations 4.28 through 4.30 show that the difference is positive.

$$\begin{aligned}
\bar{y}_C^{e,PD} &= \frac{2}{n}\Bigg\{(1-\eta)^2\Big[y(C,C)\Big] + \eta\,(1-\eta)\Big[y(D,C)\Big] + (1-\eta)\,\eta\Big[y(C,D)-\beta_0\Big] \\
&\qquad + \eta^2\Big[y(D,D)-\beta_0\Big]\Bigg\} + \frac{n-2}{n}\Bigg\{-2\eta\,(1-\eta)\beta_0 - 2\eta^2\,\beta_0\Bigg\} \qquad (4.28)\\
\bar{y}_D^{e,PD} &= \frac{2}{n}\Bigg\{(1-\eta)^2\Big[y(D,C)\Big] + \eta\,(1-\eta)\Big[y(C,C)\Big] + (1-\eta)\,\eta\Big[y(D,D)-\beta_0\Big] \\
&\qquad + \eta^2\Big[y(C,D)-\beta_0\Big]\Bigg\} + \frac{n-2}{n}\Bigg\{-2\eta\,(1-\eta)\beta_0 - 2\eta^2\,\beta_0\Bigg\} \qquad (4.29)\\
\Delta_{D,C} &= \bar{y}_D^{e,PD} - \bar{y}_C^{e,PD} \\
&= \frac{2}{n}\Bigg\{\Big[\eta\,(1-\eta) - (1-\eta)^2\Big]y(C,C) + \Big[(1-\eta)^2 - \eta\,(1-\eta)\Big]y(D,C) \\
&\qquad \Big[\eta^2 - \eta\,(1-\eta)\Big]y(C,D) + \Big[(1-\eta)\,\eta - \eta^2\Big]y(D,D)\Bigg\} \\
&= \frac{2}{n}\Bigg\{(1-\eta)(1-2\eta)\Big[y(D,C)-y(C,C)\Big] + \eta\,(1-2\eta)\Big[y(D,D)-y(C,D)\Big]\Bigg\} \\
\Delta_{D,C} &= \frac{2}{n}\Bigg\{\underbrace{(1-\eta)}_{>0}\underbrace{(1-2\eta)}_{>0}\underbrace{\Big[1+a-1\Big]}_{>0} + \eta\,\underbrace{(1-2\eta)}_{>0}\underbrace{\Big[b-0\Big]}_{>0}\Bigg\} > 0 \qquad (4.30)
\end{aligned}$$

Thus, if condition 1 and 2b of lemma 3 hold simultaneously, the N-types in the group always cooperate and a mutant of type PD would have an advantage if he played cooperation in Γ as well. By defecting, he increases his advantage over the N-types even further. He will, therefore, generate the highest payoffs in the group and reproduce twice in t_0+1, increasing the share of PD-types in the population.

In summary of the argument above, a mutant will reproduce twice if condition 1 and either condition 2a or 2b hold. ■

Remark. If neither condition 2a nor 2b hold, survival of the initial mutant depends on the relative size of the parameter a and b in Γ. In the argument below I will disregard the small level of noise, i.e., assume $\eta = 0$, for the sake of simplicity.[21]

[21] Of course, the argument remains valid for $0 < \eta < \frac{1}{2}$.

Case 2a:
a $> \alpha \geq$ **b**

	C	**D**
C	1, 1	**0,** $>$**1**
D	$>$1, 0	$\leq 0, > 0$

N vs. PD

	C	**D**
C	1, 1	**0,** $>$**1**
D	$>$**1, 0**	$\leq 0, \leq 0$

N vs. N

Case 2b:
b $> \alpha \geq$ **a**

	C	**D**
C	1, 1	$0, >1$
D	$\leq$1, 0	$>$**0,** $>$**0**

N vs. PD

	C	**D**
C	**1, 1**	$0, \leq 1$
D	$\leq$1, 0	$>0, >0$

N vs. N

Figure 4.4: Payoff in Γ^e if one mutant of type PD enters a group of N-types - cooperation is not a dominant strategy.

In case 2a shown in figure 4.4, the advantage of defection if the other player cooperates, a, exceeds b, the advantage of defection if the other player defects. In this case, the N-type cooperates if paired with the mutant, who always defects. If two N-types play against each other, there are two possible equilibria with the need for a coordination mechanism to select either one. In both equilibria one of the N-types receives $1 + a > 1$, while the other player gets nothing. The mutant's payoff in Γ always equals $1 + a > 1$, while the N-type can only receive $1 + a$ if playing against another N-type, who choses to cooperate. If an N-type encounters the mutant, he receives 0. Thus, the average payoff of the mutant in Γ is strictly bigger than the payoff of the non-mutant, and the mutant has an advantage in Γ^e if inequality 1 holds and punishing is relatively advantageous for the mutant.

The second case in figure 4.4 shows the situation in which the N-types could have a payoff advantage. In this case, N-types chose to cooperate among themselves but to defect against the PD-type mutant. The mutant receives a payoff of b every time he is selected to play Γ and nothing otherwise. The N-type receives $b - \alpha < b$ in Γ^e when playing against the mutant. However, this is the case only in 1 out of the $n - 1$ cases that the N-type is selected to play Γ. In $n-2$ out of $n-1$ cases the N-types generate a payoff of $1 > b$. Which of the two effects dominate depends on the relative size of a, b, α and group size n_{t_0}. □

Let the maximum group size that fulfils inequality 4.27 be denoted by

$$\underline{n} = \max \left\{ n \;\middle|\; n < \frac{\alpha}{\beta_0} + 1,\ n \in \mathbf{N} \right\} \tag{4.31}$$

Next, the question of fixation is discussed.

Lemma 4 (Fixation of types). *Assume no changes in group size and no additional mutations for the next T_{fix} periods after t_0. If a mutant of type PC or PA successfully reproduces after the initial period t_0, fixation happens, i.e., all players in the group will be of the mutant's type, after T_{fix} periods, where*

$$T_{fix} = \{T \mid T \geq \log_2 n_{t_0};\ T \in \mathbf{N}\} \tag{4.32}$$

The above is also true for successful mutants of type PD that either

1. *punish defection so strongly to enforce cooperation right after the initial mutation in t_0, i.e., $\alpha \geq \max\{a, b\}$* **or**

2. *punish defection only so weakly that all but one group members needs to become PD-types and punish defection collectively to change any group member's behaviour to cooperation, or cooperation cannot be enforced in the group at all i.e., $(n_{t_0} - 1)\ \alpha < min\{a, b\}$*

Proof. As long as the conditions for proposition 1 (relative fitness in small groups) are fulfilled, genetically determined punishers have an advantage over the N-types, replacing more and more of them, until all group members are of the mutant's type. The two conditions will be verified one by one.

Condition 1 (sufficiently small group): By assumption, the initial mutant reproduced successfully. This implies that the group is sufficiently small for inequality 4.19 in proposition 1 to hold.

Condition 2 (same strategy in Γ): With one exception, all players select the same strategy in each period $t_1 = t_0 + 1$ through $t_{fix} = t_0 + T_{fix}$.

If only players of type PA and/or PC enter a group of N-types, corollary 1 on page 94 implies that no group member will ever select cooperation in Γ. Thus, all group members will play the same strategy in Γ, which is defection.

For PD-types the selection of strategies in Γ depend on parameters α, a and b in a similar way to what has been discussed in the proof of lemma 3. In particular, situations are possible where N-types cooperate with each other while defecting against PD-types. Conditions 1 or 2 prevent this situation.

If PD-types punish defection strong enough, all players are punished for defection by at least $\alpha > \max\{a, b\}$. This includes the mutant PD-types themselves because after period t_0 there are two or more PD-types in the group. By lemma 2, all players will select cooperation.

If PD-types punish defection only weakly, i.e., $\alpha < \frac{\min\{a,b\}}{n_{t_0}-1}$, then all players will play defection unless all but one player in the group is of type PD. Thus, in all but the last period before fixation, the PD-types have a relative advantage according to proposition 1.

In the last period before fixation, the remaining N-type will select cooperation in Γ, but always play against a PD-type, who defects. In this way, PD-types generate higher payoffs than the remaining N-type in Γ. As the PD-type would have an advantage when selecting the same strategy in Γ and can generate a higher payoff by defecting, PD-types have a relative advantage over the remaining N type also in the last period before fixation.

Thus, fixation will happen under the conditions given in this proposition. The number of periods from initial mutation to fixation is given by

$$T_{fix} = min\left\{T \mid 2^T \geq n_{t_0}\right\} \tag{4.33}$$

which can easily be rearranged to equation 4.32. ■

Finally, the above results for sufficiently small groups ($n_{t_0} \leq \underline{n}$) can be extended to type compositions other than groups of only N-types. Lemma 3 and 4 establish that mutants of type PA, PC, and, under certain conditions, PD reproduce successfully and their descendants take over the small group, eliminating the N-types completely.

Corollary 2 provides a ranking among the different types of genetically determined punishers in groups smaller than $\underline{n}$, depending on the strategy in Γ that players select. The PA-types, who punish other group members irrespective of their actions in Γ, will always dominate, irrespective of group composition and strategy selected in Γ.

PD-types dominate N- and PC-types as long as defection is played. This is the case for weak punishers of defection ($(n_{t_0} - 1)\ \alpha < min\{a, b\}$). If there is one or more players of type PC in the group, they provide an additional disincentive against cooperation and prevent any change in behaviour from happening. Thus, defection will remain the only strategy selected until fixation of the PD-types.

For strong punishers of defection ($\alpha \geq \max\{a, b\}$), however, there are small groups with one or more individuals of type PC that cannot be taken over completely. If the group has at least 4 members, evolutionary dynamics can result in a co-existence of PD- and PC-types. There are two possible reasons. The first is similar to what has been discussed related to PD-types with intermediate strength of punishment ($\min\{a, b\} \leq \alpha < \max\{a, b\}$). Equilibria can occur with PC types cooperating

among each other, but defecting when playing against PD types. The situation is similar to the equilibria shown in case 2b in figure 4.4, but for PC instead of N-types. The second reason is that, as soon as the number of PD-types in the group exceeds the number of PC-types by two, all players, including the PD-types, will play cooperation. But under cooperation, the PC-types have an advantage over PD-types and, therefore will grow in numbers. The following example will show that this can result in permanent co-existence of PD and PC types.

Example. Consider a small group with $n_{t_0} = 6$, at least 3 individuals of type PC, up to 3 N-types and no individual of type PA. Assume that $\frac{\alpha}{\beta_0} > 5$, so that condition 1 for lemma 3 is fulfilled.

Now consider a mutant with type PD and $\alpha \geq \max\{a, b\}$ to enter the group. According to lemma 3 and corollary 2, the mutant will generate the highest average payoff in period 0 and reproduce twice. One of the N-types, or, if there are none, one of the PC-types will not reproduce.

Therefore, in period 1 there are 2 PD-types in the group. But because there are still at least 3 PC-types in the group as well, the threat of punishment of defection is counterbalanced by the PC-types' threat to punish cooperation and no change in behaviour occurs. All players continue playing defection. Therefore, the PD-types still have a relative advantage. Each PD-type reproduces twice, replacing PC- and N-type(s), so that in period 2 there are exactly 4 individuals of type PD and 2 of type PC in the group.

But now every player, including the PD-types face more individuals punishing defection than individuals punishing cooperation. Therefore, and because the punishment technology is characterized by an α high enough for just one player to enforce cooperation, all players will cooperate in period 2. Under cooperation, however, the PC-types have an advantage over players of type PD. Therefore, PC-types will reproduce twice after period 2 and the group in period 3 will consist of 4 PC-types and only 2 PD-types.

But this is exactly the number of PD-types as in period 1. Thus, the group experiences a cyclical fluctuation of types, with either 2 PD- and 4 PC-types or vice versa. □

The same situation can also occur for a PC-type mutant entering a group of high-α PD-types. If all players cooperate in a sufficiently small group, the PC has a relative advantage by punishing cooperation and survives the initial period. However, there cannot be fixation of type PC given this initial situation. The reason is that, under defection, the PC-types have the smallest advantage over the N-types among the 3 genetically determined punishers. Thus, if the number of PD-types fall below the number of PC-types and players start playing defection in Γ, PD-types gain an advantage over PC types and their share in the group grows again. Fixation of PC-types is only possible in groups of initially only N-types.

Summing up the results on mutation and fixation: In sufficiently small groups ($n_{t_0} \leq \underline{n}$), mutants of type PA always survive and replace players of any other type completely in finite time. PC-types survive the initial mutation in groups of only N-types or in groups with sufficiently many PD-types so that cooperation is enforced. However, fixation of type PC only happens in initially homogeneous groups of N-types.

Mutants of type PD survive the initial mutation in groups with any combination of N- and PC-types, if the punishment technology is characterized by a low intensity of punishment. If it is so low, that even $n_{t_0} - 1$ punishers of defection cannot enforce cooperation, fixation happens for sure, even in mixed groups of N- and PC-types. Very high-α punishment technology causes PD-types to survive the initial mutation in these groups as well, but allows fixation only in initially pure N-type groups. In groups with PC-types, high-α punishment can lead to permanent co-existence of PD and PC types. Any number of PA-types in the group prevent the success of PD-type mutants.

The next proposition presents the last important result for this subsection - permanent hypocrisy on the way to enforcing full cooperation. Of course, only punishers of defection can enforce cooperation. In general, playing a strategy s in the moral code of a player can only be prevented if threatened or actual punishment overcompensates the payoff advantage in Γ of playing strategy s. However, as those who have s in their moral codes punish *other* players, who play s in Γ, but not themselves, players with s in their moral codes are the last ones that are deterred from playing s.

Proposition 3 (Permanent hypocrisy). *Among all four types of players, PD-types, who only have defection in their moral code, and punish only those who are not cooperating, will stop defecting* last.

Proof. Suppose not. Then there exists a period t, in which a PD type cooperates while at least one other player finds it optimal to defect. For this period t, lemma 2 and the related corollary 1 translate into:

$$\exists\, j \text{ such that } n^{PD}_{\setminus i,\,t} - n^{PC}_{\setminus i,\,t} \geq \frac{\max\{a,b\}}{\alpha} \quad \text{and} \quad n^{PD}_{\setminus j,\,t} - n^{PC}_{\setminus j,\,t} < \frac{\min\{a,b\}}{\alpha}, \text{ with}$$
$$\sigma_i = PD \quad \text{and} \quad \sigma_j \in \{N, PA, PC\}$$

But how much stronger or weaker defection is punished compared to cooperation *depends on type* σ_i in the following way:

$$\Delta^{\sigma_i}_t \equiv n^{PD}_{\setminus i,\,t} - n^{PC}_{\setminus i,\,t} = n^{PD}_t - n^{PC}_t - \mathbf{1}_{PD} + \mathbf{1}_{PC}$$

Therefore

$$\begin{aligned}\Delta_t^{PC} &= n_t^{PD} - n_t^{PC} + 1,\\ \Delta_t^{N} &= \Delta_t^{PA} = n_t^{PD} - n_t^{PC}, \text{ and}\\ \Delta_t^{PD} &= n_t^{PD} - n_t^{PC} - 1\end{aligned}$$

But, by assumption, the differences must fulfil

$$\Delta_t^{PD} \geq \frac{\max\{a,b\}}{\alpha} \geq \frac{\min\{a,b\}}{\alpha} > \min(\Delta_t^N, \Delta_t^{PA}, \Delta_t^{PC})$$

Thus, it is required that

$$\Delta_t^{PD} > \Delta_t^N \quad \Leftrightarrow \quad n_t^{PD} - n_t^{PC} - 1 > n_t^{PD} - n_t^{PC} \quad \Leftrightarrow \quad -1 > 0 \tag{4.34}$$

which is a contradiction. ■

To underline it again: Proposition 3 is another consequence of the fact that no player punishes himself. In this model, a player's moral only affects his behaviour against other players. His moral code has no effect on his own behaviour in Γ. Players have no remorse and will only behave in a way they consider morally right, if enough others share their sentiments and (threaten to) punish playing an action in their moral code in the same way as they themselves (threaten to) punish others.

4.4.2 Evolutionary stability

This subsection discusses long term evolutionary dynamics, that allows for repeated mutations. It maintains the assumption that mutation rates are very low. In particular, simultaneous mutations re disregarded. In addition, external shocks (intergroup contact and/or changes to the quality of the environment) are not regarded here. They will be discussed in the next subsection.

Thus, the question discussed in this subsection is: What happens to a group of players that is homogeneous in type if a single mutant of a different type is introduced? I consider a type σ_i as evolutionary stable if a mutant of any other type $\sigma_j \neq \sigma_i$ does not reproduce after the initial period t_0 in a group of only σ_i-types.

Evolutionary stability depends on group size n_{t_0} relative to two parameters of the model - $\underline{n}$, the maximum group size for a genetically determined punisher to have a relative advantage as a mutant in a group of N-types, and n_c^P the minimum group size for the scale effect of punishment to kick in and reduce individual cost of punishment to ε. Assume that the efficiency of possible punishment technologies is limited and the number of individuals needed for the scale effect of punishment to apply is large,

so that

$$0 < \underline{n} < n_c^P < n_{max} \tag{4.35}$$

Evolutionary stability is discussed for three different group sizes n_{t_0}: Sufficiently small groups ($0 < n_{t_0} \leq \underline{n}$), which have been discussed extensively in the last subsection, medium sized groups ($\underline{n} < n_{t_0} < n_c^P$), and big groups with $n_c^P \leq n_{t_0} \leq \min\{N_{max}, n_{max}\}$. Of course, medium sized and large groups can only occur if the carrying capacity of the environment N_{max} is sufficiently big.

Small groups: By corollary 2, in sufficiently small groups ($n_t \leq \underline{n}$), a mutant of PA type has the biggest relative advantage over the N types, followed by the PD and the PC type. Thus, as long as group size is not bigger than $\underline{n}$, a homogeneous group of PA types is stable against mutants of any other type. By the same token, sufficiently small groups of PD, PC, or N types are not evolutionary stable, as they are vulnerable to an invasion by PA types.

Medium sized groups: If $\underline{n} < n_t < n_c^P$, i.e., the group is of medium size and there is no scale effect reducing the cost of punishment to ε, then genetically determined punishers have a relative disadvantage. Thus, medium size groups of N types are stable against mutants, and none of the genetically determined punishers can resists an invasion of N types.

Large groups: Proposition 4 establishes the result for large groups.

Proposition 4. *Consider large groups of individuals, i.e., with $n_{t_0} \geq n_c^P$, that is homogeneous in type. In these groups, all types are evolutionary stable.*

Proof. For N types proposition 4 is obvious. N types have an advantage, because with a single genetically determined mutant there is no scale effect to sufficiently reduce the cost of punishment in a big group. Therefore, $\beta = \beta_0$. By assumption $n_{t_0} \geq n_c^P$ and, according to inequality 4.35, $n_c^P > \underline{n}$. Thus, $n_{t_0} > \underline{n}$ and a mutant that is genetically determined to punish any strategy $s \in S$ has a disadvantage relative to N-types.

In a homogeneous group of genetically determined punishers (PD-, PC- or PA-types) punishment happens irrespective of the type of the punishers because of strictly positive noise in behaviour. But in groups with size $n_t > n_c^P$, the cost of punishment for strategies in the moral code of the group members is ε. Because $\beta = \varepsilon < \frac{\alpha}{n_{max}}$ by assumption, $n_t \leq n_{max} < \frac{\alpha}{\varepsilon} + 1 = \frac{\alpha}{\beta} + 1$ and punishers have a relative advantage in average payoff according to proposition 1. Therefore, such a group is evolutionary stable against N type mutants.

But the group also resists invasions of other types of genetically determined punishers. If the group consists of PA types, the PA types have an advantage over mutants of PD or

PC types for the same reason as in very small groups: Any kind of punishment has cost of ε and is, therefore, advantageous. And if every kind of punishment is advantageous, the PA type, who punishes most, has the biggest advantage.

A group of only PD types is stable against a single mutant of PA or PC type because the mutant will be the only one punishing cooperation, not enjoying any scale effect and, therefore, generating a lower payoff than the PD type who does. The same is true for a mutant of PD or PA type in a group of PC types.

Thus, large groups of genetically determined punishers are stable against a single mutant of type N and against invasions by the two other types of genetically determined punishers, too. ■

Let me sum up the results for evolutionary stability in homogeneous groups of different sizes: For very small groups with $n_{t_0} \leq \underline{n}$, only groups of unconditional punishers (PA-types) are evolutionary stable. Medium sized groups with $\underline{n} < n_{t_0} < n_c^P$ are only evolutionary stable if they consist of N types, who are not genetically determined to punish. Big, homogeneous groups with $n_{t_0} \geq n_c^P$ are always stable against single mutants of any other type, irrespective of the group members' type.

4.4.3 Selection of moral codes

The previous subsection took an inward looking perspective on stability of small groups and discussed evolutionary stability against single mutations. This subsection deals with external shocks to a group. Again, the analysis focuses on groups that are homogeneous in type.[22]

For this final part of the discussion, some restrictions are imposed on the model's parameters to avoid further case by case discussions. Of course, the set of simultaneously admissible parameters remains non-empty and the dynamics investigated in this sub-section happen with strictly positive probability in the unrestricted parameter space.

$N^L = \underline{n} < \frac{n_{max}}{2}$**.** A bad environment can only sustain very small groups.

$\underline{n} < n_c^P \leq n_{max} - \underline{n}$**.** The threshold number of coordinated punishers needed for individual cost of punishment to become insignificantly small is smaller than the maximum theoretical group size, but bigger than maximum group size in the

[22] This simplification seems prudent given that fixation of any given type takes very few peiods only (see lemma 4, equation 4.32). The results also apply to mixed groups of PD- and PC-types, the only relevant case of sustainable heterogeneous groups.

bad environment. This implies that cost of punishment $\beta_i = \beta_0$ for all players i in the bad state of environment. It can only fall to ε in the good state.

$\frac{\max\{a,b\}}{n_{max}-\underline{n}} \leq \alpha < \frac{\min\{a,b\}}{\underline{n}-1}$**.** The punishment technology is such that cooperation cannot be enforced in very small groups ($n_t \leq \underline{n}$). Enforcement of a moral code is a pure social phenomenon.[23] But the critical number of punishers of defection in excess of punishers of cooperation is also smaller than the maximum theoretical group size n_{max}. This way, cooperation can be enforced in groups that are not perfectly homogeneous of type PD.

$0 < \eta < \frac{1-b}{n_{max}+3}$**.** The level of noise is strictly positive, but very small.

Two environmental scenarios are considered:

In scenario one small groups live in a stable environment, i.e., $f_{env} = 0$. An example closely related to human ancestors can be seen in Bonobos (Pan paniscus). The Bonobos' habitats are situated in the humid forests between the rivers Congo and Kasai in central Africa (Caswell, Mallick, Richter, Neubauer, Schirmer, Gnerre & Reich 2008). This environment has been stable for a very long time and, due to the primates' inability to swim across the rivers, the Bonobos were confined to their habitats.

The second scenario exhibits environmental change, i.e., $f_{env} > 0$. These changes can happen if a small group migrates beyond its current habitat and/or because rapid climate fluctuations that affects the current habitat of the group. Both kinds of changes happened repeatedly during human evolution, after our common ancestors left the jungle, entered the open woodlands of the African savanna, and finally spread across the globe.

As assumed, there is frequent intergroup contact in both scenarios, i.e., f_{int} is strictly positive.

First stable environments are discussed. Proposition 5 shows that a stable habitat does not allow for the evolution of moral punishment, irrespective of the quality of the environment or the action that is punished.

Proposition 5 (No success for punishers in stable environments). *In a stable environment, i.e., any habitat with $f_{env} = 0$, only groups of N-types can exist for more than $\frac{1}{f_{int}}$ periods.*

Proof. A stable habitat can either be a good or a bad one. It is populated initially by only N-types. It must be shown that genetically determined punishers do not develop or, if

[23] This loads the dice against the evolution of cooperation.

they do, they don't persist for longer than the next intergroup contact in either of the two environments.

Bad environment ($N_{max} = N^L$): In a bad environment, no group can be bigger than N^L. By assumption, $N^L = \underline{n}$. Thus $n_t \leq \underline{n}$, $\forall t$ and all groups are sufficiently small for genetically determined punishers to reproduce and eventually even fix their types in the group. Actual punishment starts to happen as soon as the first mutant of type PA, PD, or PC enters a group.

But because, by assumption, cooperation cannot be enforced in a very small group the impact and cost of punishment is not offset by higher payoffs in Γ. Therefore, in groups with genetically determined punishers, the average payoff will be lower than in groups of N-types, i.e., $\bar{\pi}^G < b$ for all groups with one or more genetically determined punishers. As assumed in equations 4.7 and 4.8, a group goes extinct in the event of intergroup contact, if the average payoff in the original group is lower than in the group it encounters. But, as mutations are very rare events, groups with one or more genetically determined pushers are surrounded by groups of N-types. In groups of only N-types the average payoff is b. Therefore, a group with one of more genetically determined punishers gets eradicated the first time it encounters another group.

Good environment ($N_{max} = N^H$): In a good environment, the situation is the same for groups with any number of PA- and PC-types, but too few PD-types to enforce cooperation. All players continue to select defection and have to bear the impact and cost of punishment. Therefore those groups cannot exist for longer than the next contact to a group of only N-types.

If there is a sufficient number of PD-types in a big group, cooperation can be enforced and the cost of punishment is reduced to ε. Thus, big groups in a good habitat with $\bar{\pi}^G \geq b$ are theoretically possible. And, according to proposition 4, they are evolutionary stable. However, in a world without genetically determined punishment, a group of PD-types can only develop after a successful initial mutation. And the initial mutant needs a small group to be successful. Big groups of N-types are evolutionary stable. And even if there was a small enough group in a good habitat, it will not grow, as growth in group size only happens if the state of the environment *changes* from bad to good. Thus, any group of PD-types will remain small, even in a good habitat. It will go extinct with the next contact to one of the groups of only N-types surrounding it.

Conclusion: In both bad and good environment, no group of genetically determined punishers can develop that generates an average payoff exceeding b, the payoff in the surrounding groups of only N-types. Thus, intergroup contact will eradicate any group with one or more genetically determined punishers. Intergroup contact happens with frequency f_{int}. Therefore, the maximum duration of existence for a group with even one genetically determined punisher in a stable environment is $\frac{1}{f_{int}}$. ■

Finally, this chapter analyzes the evolution of moral punishment under the assumption of volatility in environmental quality. It shows that environments in which the environment frequently changes in quality will eventually produce large groups of punishers of defection. In these groups, cooperation is enforced and, in the event of intergroup contact, PD-types will replace all other types. This way, the type PD is fixed in the environment. If the environment is populated by PD-types only, the population as a whole is evolutionary stable against mutants of all other types.

For all players in a population to become PD-types, three conditions have to be fulfilled with strictly positive probability.

1. A mutant of type PD reproduces after the initial period.
2. PD-types fix their type within their group.
3. PD-types fix their type in the entire population. This means that all groups become groups of only punishers of defection.

The above conditions need to be fulfilled with strictly positive probability. If they are, a population of only PD types will eventually appear, as time is assumed to be infinite. However, the population also needs to stay pure PD-type for punishment of defection to be a universal and persistent feature of human nature. Therefore, a last condition has to hold with certainty.

4. A population of only PD-types is stable against mutants of any other type.

The 4 conditions are investigated one by one.

Lemma 5 (Success of a PD-type mutant). *In an environment with changing quality, populated by groups of N-types, a mutant of type PD reproduces with strictly positive probability.*

Proof. A volatile environment has two states, a good and a bad one. According to proposition 1, an initial mutation in group of only N-types is successful in sufficiently small groups. By assumption, no group is bigger than $\underline{n}$ if the environment is in the bad state. Thus, initial mutations in groups of N-types are successful if the environment is in the bad state. As the two states are alternating with a constant frequency f_{env}, a mutation is happening in a bad environment with positive probability $\left(\frac{1}{2}\right)$. ■

Lemma 6 (Fixation of type PD in a group). *In an environment with changing quality, mutants of type PD will fix their type within their group with strictly positive probability.*

Proof. In a sufficiently small group in isolation, fixation of type happens in finite time, as shown in lemma 4. However, a small group of PD-types is neither stable against mutants of type PA nor against intergroup contact with groups of N- or PC-types. However, mutations are rare and intergroup contact does not happen in every period. Thus, with positive probability, a change in the environment happens before a small group of PD-types can be eradicated due to mutation or intergroup contact.

If there exists a group with any number of PD and N-types in the population, when the environment changes from bad to good state, all players will be of type PD no later than one period after the environmental shock. To see this, consider the type composition before the environmental shock at t_{env}.

$$n_{t_{env}-1} \leq \underline{n}; \qquad 0 < n^{PD}_{t_{env}-1} \leq \underline{n}; \qquad n^{N}_{t_{env}-1} = n_{t_{env}-1} - n^{PD}_{t_{env}-1} < \underline{n} \tag{4.36}$$

According to proposition 1, any mutant of type PD generates higher payoffs than N-types in a sufficiently small group. Therefore, when the group grows to n_{max} under the good environment, all additional players will be of type PD. Thus:

$$n_{t_{env}} = n_{max}; \quad n^{N}_{t_{env}} = n^{N}_{t_{env}-1}; \quad n^{PD}_{t_{env}} = n_{max} - n^{N}_{t_{env}} > n_{max} - \underline{n} \geq n^{P}_{c} \tag{4.37}$$

As $n^{PD}_{t_{env}} \geq n^{P}_{c}$, the cost of punishment β is reduced to ε. Therefore, PD-types retain their relative advantage also in the big group. At the same time $n^{PD}_{t_{env}} > n_{max} - \underline{n} > \frac{n_{t_{env}}}{2}$. As more than half the group members are of type PD, the remaining N-types are replaced by PD-types after the first period in the big group. ■

Lemma 7 (Fixation of type PD in the population). *In an environment with changing quality, mutants of type PD will fix their type within the entire population with strictly positive probability.*

Proof. The dominance of PD types in the population rests on the fact that only within large groups of PD types cooperation is enforced. A higher average payoff is achieved in those groups and they eliminate groups of other types.

The proof follows this line of argument. First it needs to be shown that, as a group with any number of PD-types experiences a positive shock to its environment, all players in the group start to cooperate.

According to lemma 2, a player i cooperates if

$$n^{PD}_{\setminus i} - n^{PC}_{\setminus i} \geq \frac{\max\{a, b\}}{\alpha}$$

This condition is fulfilled for all players in a group of only PD- and N-types after the transition to the good state of the environment. To see this, consider the following. By assumption,

$$\alpha \geq \frac{\max\{a, b\}}{n_{max} - \underline{n}}; \quad n^{PD}_{t_{env}} > n_{max} - \underline{n} \quad \text{and} \quad n^{PC}_{\setminus i} = 0, \ \forall i \tag{4.38}$$

therefore

$$n_{t_{env}}^{PD} \geq n_{max} - \underline{n} + 1 \quad \text{and} \quad n_{t_{env}}^{PD} - 1 \geq \frac{\max\{a,b\}}{\alpha} \tag{4.39}$$

But

$$n_{\backslash i,\, t_{env}}^{PD} = \begin{cases} n_{t_{env}}^{PD} & \text{for } N\text{-types} \\ \\ n_{t_{env}}^{PD} - 1 & \text{for } PD\text{-types} \end{cases} \tag{4.40}$$

and therefore

$$n_{\backslash i,\, t_{env}}^{PD} - n_{\backslash i,\, t_{env}}^{PC} = n_{\backslash i,\, t_{env}}^{PD} \geq n_{t_{env}}^{PD} - 1 \geq \frac{\max\{a,b\}}{\alpha};\ \forall i \tag{4.41}$$

and cooperation is enforced for all players in the group.

To demonstrate that groups enforcing cooperation in the good environment will eliminate others in the event of intergroup contact, the average payoff to their members needs to be compared to the average payoff in all other groups. Let homogeneous groups of N-types, where defection is played but no punishment happens, be the baseline again.

All groups with genetically determined punishers, but a sub-critical number of PD-types to enforce cooperation will generate lower average payoffs than a homogeneous groups of N-types, because they do not generate higher payoff in Γ but bear the impact and cost of punishment. Average payoff across all group members in a homogeneous group of N-types is

$$\bar{\bar{y}}_t^{e,\,N} = \frac{2}{n_t}\Big[(1-\eta)^2\, b + \eta\,(1-\eta)\, 0 + (1-\eta)\,\eta\,(1+a) + \eta^2\, 1\Big] \tag{4.42}$$

In groups enforcing cooperation, defection only happens by accident. Therefore, for sufficiently small amounts of noise η, large groups enforcing cooperation generate higher average payoffs than those groups in the population that continue playing defection. For a group of N- and PD-types, with all players selecting cooperation in Γ, average payoff is

$$\begin{aligned} \bar{\bar{y}}_t^{e,\,\{N,PD\}} \;=\; & \frac{n_t^{PD}}{n_t}\Bigg\{\frac{2}{n_t}\Big[(1-\eta)^2\, 1 + \eta\,(1-\eta)(1+a-n_{\backslash i,\,t}^{PD}\alpha) - (1-\eta)\,\eta\,\beta \\ & + \eta^2(b - n_{\backslash i,\,t}^{PD}\alpha - \beta)\Big] + \frac{n_t-2}{n_t}\Big[(1-\eta)^2\, 0 - 2\,\eta(1-\eta)\beta - \eta^2\, 2\beta\Big]\Bigg\} \\ & + \frac{n_t - n_t^{PD}}{n_t}\Bigg\{\frac{2}{n_t}\Big[(1-\eta)^2\, 1 + \eta\,(1-\eta)(1+a-n_{\backslash i,\,t}^{PD}\alpha) \\ & - (1-\eta)\,\eta\, 0 + \eta^2(b - n_{\backslash i,\,t}^{PD}\alpha)\Big] + \frac{n_t-2}{n_t}\, 0\Bigg\} \end{aligned} \tag{4.43}$$

The first term in curly brackets is the payoff for PD-types, averaged over those who play Γ and those who only act as punishers. The second term is the payoff for each N-type that is still in the group.

Let $\Delta_{\bar{\bar{y}}_t^e} = \bar{\bar{y}}_t^{e,\{N,PD\}} - \bar{\bar{y}}_t^{e,N}$ indicate the advantage in average payoff in groups enforcing cooperation over groups of only N-types, who defect, but never punish.

$$\begin{aligned}
\Delta_{\bar{\bar{y}}_t^e} &= \frac{n_t^{PD}}{n_t}\Bigg\{\frac{2}{n_t}\Big[(1-\eta)+\eta\,(1-\eta)(a-[n_t^{PD}-1]\,\alpha-\beta)+\eta^2(b-[n_t^{PD}-1]\,\alpha-\beta)\Big] \\
&\quad -\frac{n_t-2}{n_t}\,2\eta\beta\Bigg\}+\frac{n_t-n_t^{PD}}{n_t}\frac{2}{n_t}\Big[(1-\eta)+\eta(1-\eta)(a-n_t^{PD}\,\alpha)+\eta^2(b-n_t^{PD}\,\alpha)\Big] \\
&\quad -\frac{2}{n_t}\Big[(1-\eta)^2\,b+(1-\eta)\eta(1+a)+\eta^2\Big] \\
&= \frac{2}{n_t}\Bigg[\Big\{(1-\eta)+\eta\Big((1-\eta)\,a+\eta\,b-n_t^{PD}\,\alpha\Big)+\frac{n_t^{PD}}{n_t}\Big(\eta\,(\alpha-(n_t-1)\beta)\Big)\Big\} \\
&\quad -\Big\{b-2\eta\,b+\eta^2 b+(1-\eta)\eta+\eta(1-\eta)\,a+\eta^2\Big\}\Bigg] \\
&= \frac{2}{n_t}\Bigg[(1-\eta)^2-\eta^2-\eta\,n_t^{PD}\alpha-(1-2\eta)\,b+\frac{n_t^{PD}}{n_t}\Big(\eta\,(\alpha-(n_t-1)\beta)\Big)\Bigg] \\
&= \frac{2}{n_t}\Bigg[(1-b)(1-2\eta)+\eta\,n_t^{PD}\left(\frac{1}{n_t}(\alpha-(n_t-1)\,\beta)-\alpha\right)\Bigg] \\
&= \frac{2}{n_t}\Bigg[(1-b)-\eta\left(2(1-b)+\frac{n_t-1}{n_t}n_t^{PD}(\alpha+\beta)\right)\Bigg] \qquad (4.44)
\end{aligned}$$

The term in square brackets is positive for sufficiently small level of noise η:

$$\Delta_{\bar{\bar{y}}_t^e} > 0 \quad \Leftrightarrow \quad \eta < \frac{1-b}{2(1-b)+\frac{n_t-1}{n_t}n_t^{PD}(\alpha+\beta)} \qquad (4.45)$$

But $n_t = n_{max} > n_c^P$, and, therefore, cost of punishment $\beta = \varepsilon < \frac{\alpha}{n_{max}}$. Taking the restrictions on all parameters into consideration, a lower bound for the maximum level of noise for groups enforcing cooperation can be derived:

$$\eta < \frac{1-b}{\underbrace{2(1-b)}_{\leq 2}+\underbrace{\frac{n_t-1}{n_t}}_{\leq 1}\underbrace{\frac{n_t^{PD}}{n_{max}}}_{\leq 1}(n_{max}+1)\,\alpha} \qquad (4.46)$$

Because

$$\alpha < \frac{\min\{a,b\}}{\underline{n}-1} < 1 \qquad (4.47)$$

the assumption on the maximum level of noise

$$\eta < \frac{1-b}{2+n_{max}+1} \tag{4.48}$$

implies that $\Delta_{\bar{\bar{y}}_t^e} > 0$ and big groups enforcing cooperation have an advantage over groups playing defection, even if no punishment happens in those groups as they consist of only N-types.

Inequality 4.48 shows that, on the one hand, enforcing cooperation is the more advantageous over playing defection without punishment the bigger is the difference between payoff in the cooperation and the defection equilibrium in Γ. On the other hand, the advantage is the smaller the bigger is the group and the more punishment happens every time defection is played by accident. The extent of defection and punishment is mitigated, however, by a very low level of noise. Therefore, groups enforcing cooperation only have an advantage over non-punishing N-type groups for sufficiently low levels of behavioural noise. ■

Taking lemma 5, 6 and 7 together, the above showed that, with strictly positive probability, a population of only PD-types evolves. Finally it needs to be verified that this population is evolutionary stable.

Lemma 8 (Stability of a PD-population against single mutants). *In an environment with changing quality, a population of only PD-types is stable against mutants of any other type.*

Proof. The population of PD-types needs to be stable against mutants of type N, PC and PA, in both the good ($N_{max} = N^H$) and the bad ($N_{max} = N^L$) environment.

In the good environment group size $n_t = n_{max} > n_c^P$; $\forall t$ and proposition 4 implies that all groups are evolutionary stable against single mutants. Therefore, all groups remain homogeneous PD-type groups and the entire population is stable against mutants of any type.

If the environment is in the bad state, group size $n_t \leq \underline{n}$; $\forall t$ and all players select defection. According to corollary 2, in sufficiently small groups those who punish most have the biggest relative advantage. With a small level of noise, PD-types punish more often than PC and N-types. Therefore, PD-types have a relative advantage and their groups are evolutionary stable against mutants of N and PC-type.

A mutant of type PA, however, successfully reproduces and fixes his type in a group of PD-types. But, contrary to PD-types, PA-types also punish in the event of cooperation. Therefore, for $\eta > 0$, payoff in small groups with PA-types is reduced more than in groups of only PD-types. Because a group with mutant PA-types is surrounded by homogeneous PD-type groups it will be eliminated the first time it encounters another group.

Only PD-type groups survive and the population stays homogeneous in type. ■

Taking all the above together, proposition 2 can finally be proved very easily (the proposition is restated here for convenience):

Proposition 2 (The dominance of cooperation). *Consider an environment that changes in quality between the good and the bad state, ie.* $f_{env} > 0$.

In such a volatile environment, all players in a population will eventually be punishers of defection. Mutants of other types do not reproduce. Cooperation will be the norm shared by all non-mutant players.

The only temporary exception can be PA*-types in an isolated group.*

Proof. According to lemma 5, a mutant of type PD reproduces with positive probability in the original situation of only N-types - whenever the environment is in the bad state. These states occur with a constant, positive frequency. If a mutant of type PD successfully reproduces after the initial period, lemma 6 ensures that there is a strictly positive probability that he fixes his type in his group. This happens if there is no intergroup contact and no mutant of type PA appears before all N-types are replaced.

Lemma 7 maintains that PD-types fix their type within the entire population with strictly positive probability. If the environment improves as long as a group of (predominantly) PD-types exists, the number of punishers increases to a level that enforces cooperation in the big group. For sufficiently small levels of behavioural noise μ, average payoff in these large groups choosing cooperation is higher than in the other groups choosing defection. Thus, groups of N, PC and PA-types, or mixtures thereof, are replaced when they encounter a group of cooperators.

This happens in the good state among big groups - as long as no bad environmental shock happens. As changes between environmental states only happen with finite frequency there is a positive amount of time in which the group(s) of PD-types can interact with all other groups and fixation in the population can happen. Thus, with positive probability, all individuals within the population become punishers of defection.

Because time is infinite in the model considered here, events with strictly positive probabilities will eventually occur. Thus, a given population will become purely of type PD. Finally, lemma 8 proved that a population of only PD-types is evolutionary stable against mutants of any other type. This is certain and does not depend on the state of the environment. Therefore, once a population has changed to PD-type only, it will stay this way forever. The only non-PD-types in the population at a given point of time are individuals whose type changed due to a mutation when they were born into the very period considered or PA-type individuals as long as their group is isolated against contact with the neighbouring groups of PD types that punish less and, therefore, have higher average payoffs. ■

4.5 Conclusion

This chapter discusses the evolution of morally motivated punishment in the early stages of human evolution. In a small group, continuous interaction model it shows that mutants, who are genetically determined to punish a certain strategy played in a prisoner's dilemma, can successfully reproduce and even eliminate all players of other types in the group.

Their initial success rests on their *relative* advantage in very small groups. The mutant is able to optimize his strategy in Γ in the same way as all other players. But if a punisher and a non-punisher both find it optimal to play a strategy that is punished, the punisher is punished by one player less - himself. This advantage can overcompensate the cost of punishing others, if the cost of punishing all other group members is sufficiently low. This is the case if the group, in which the initial mutation happens, is sufficiently small. When actual punishment is advantageous in a very small group, mutants that punish the *most* are most successful there.

As a consequence of the above, only the type of players who always punishes, irrespective of the actions played in the prisoner's dilemma, is evolutionary stable in very small groups. These players punish most and have an advantage over mutants of any other type in the population.

In homogeneous big groups, all types are evolutionary stable. The reason is that, by assumption, the individual cost of punishment is reduced to some positive but very low level if a large number of players punish the same strategy at the same time. This is the case for all types of genetically determined punishers, irrespective of what strategies they punish. Big groups of non-punishers are evolutionary stable, because punishing mutants find the cost of punishing all other group members too high to be compensated by the fact that they do not punish themselves.

However, a type that is evolutionary stable in its group is still not guaranteed to survive in the presence of lethal inter-group contact. While a mutant's payoff can be higher *relative to other group members*, his *absolute* payoff is lower compared to individuals in small groups without any punishment, if players in both groups play the same strategy in the prisoners dilemma. This payoff difference is even bigger for his non-punishing group members.

The model assumes that, in the event of intergroup contact, the group with lower average payoff gets eliminated. It is replaced by a group with the same type-composition than the group with the higher average payoff. Groups with genetically determined punishers that cannot enforce cooperation continue playing defection in

the same way as non-punishers. But non-punishers receive higher average payoffs, as they do not bear the cost and impact of punishment. Therefore, genetically determined punishers that cannot enforce cooperation are in general not stable against intergroup contact.

Cooperation can be enforced in a group if so many more players punish defection than cooperation that the payoff advantage of defection in the prisoners dilemma is overcompensated. The group needs to be sufficiently big for this to happen. If cooperation can be enforced, the increased payoff in the prisoners dilemma can overcompensate the impact and cost of punishing those players who fail to play the optimal strategy, given that this behavioural noise is sufficiently low. Therefore, large groups of punishers of defection can enforce cooperation and prevail in the event of intergroup contact.

But how can these big groups of punishers of defection emerge? A single mutant in a big group cannot be successful, because, by assumption, the punishment technology is not efficient enough for a single player to generate an advantage from punishing a large number of non-punishers. And extremely unlikely events like simultaneous mutations of a big number of individuals are not considered in this research.

The only way for a big group of genetically determined punishers to emerge starts with mutation and fixation of type in a small group. However, group size is assumed to be constant in this model, unless there are external shocks to the group's habitat. In a stable habitat, genetically determined punishers die right after the initial mutation, if it is a good habitat. If the environment is in a bad state and habitats only allow for small groups, groups of mutants cannot persist for longer than the next lethal contact to a group of non-punishers. The development of big groups of genetically determined punishers requires a volatile environment.

Moreover, the model shows that a volatile environment favours punishers of defection - eventually they will fix their type in the entire population. Most important for their success is that the moral behaviour they enforce is increasing payoffs in equilibrium in the prisoners dilemma. Thus moral punishment and cooperation necessarily co-evolve.

Chapter 5

Conclusion

This thesis analyzes behaviour that seems "irrational" in light of standard economic theory. It looks at altruism, discrimination and punishment of immoral behaviour at a cost, but no immediate benefit to the punisher. Both culture and evolution are investigated as potential determinants of these prominent pattern of human behaviour. Cultural influences are identified by looking at empirical differences in behaviour between groups of children that are distinguished by their language or phenotype. That evolution shaped human psychology in the past to "hardwire" certain patterns of behaviour is argued in a model of human evolution in small groups. This thesis combines empirical and theoretical work and presents both experimental evidence for socialization in a specific cultural context for sharing and discrimination and theoretical arguments that we are biological determined to punish violations of moral codes.

Experimental work with more than 1,000 children in Vancouver, BC, Canada allowed us to test a multitude of hypotheses about the drivers behind altruism of children in the form of sharing with unrelated others and discrimination based only on visibly different phenotypes of other children. Only a fraction of the rich dataset has been utilized up to now. A large number of questions can still be addressed, ranging from the influence of different school types on attitudes and behaviour of children to the influence of exposure to visibly different phenotypes on friendship patterns and racial stereotypes among children.

The existing results of the empirical work reported in chapters 2 and 3 point very strongly at culture and socialization as the shaping force behind altruism and discrimination. In fact, children with different cultural backgrounds behave significantly different when asked to share at the age of five, but loose some of the difference with extended exposure to socialization in a public Canadian school environment. With regards to discrimination, the findings are more complex. While white children

have a clear sense of identity and share more with hypothetical others of White phenotype, East Asian children do not show in-group favouritism. Instead they show signs of out-group favouritism. Again, culture seems to be a driving force. Out-group favouritism as a pattern of "irrational" behaviour cannot be identified among children with non-East-Asian family background. It seems to be very specific to the culture of immigrants from East Asia. However, the evidence is rather weak and the identification of out-group favouritism as a general feature of East Asian culture requires further research.

While the empirical work makes a strong case for culture as the driver behind "irrational" behaviour, the theoretical model shows that under volatile environmental conditions, punishment of immoral behaviour could be evolutionary advantageous for individuals in small groups. Chapter 4 argues that with sufficiently efficient punishment technologies, a mutant that is genetically determined to punish immoral behaviour in a prisoners' dilemma could thrive, and the punisher types could take over a small group of individuals. If the moral code that the punishers defend is payoff-increasing, the punisher type will finally be fixed in the entire population. Boundary conditions for the environment of human evolution are specified under which fixation of types and evolutionary stability will happen.

Do these arguments contradict each other? Not necessarily. Even though we might be biologically determined to punish immoral behaviour because nature has given us a strong psychological urge to do so, the definition of the relevant moral code can be the outcome of a cultural process. Looking back at human history shows that moral codes are in constant change. But whatever the moral code was at a given time, it has been fiercely defended most of the time.

Research in biology also supports the duality between nature and nurture as determinants of behaviour. Findings in evolutionary anthropology suggest that vengeful behaviour could have a biological basis because it can be observed also with chimpanzees, the closest relative to humans (Jensen, Call & Tomasello 2007*b*). Concerns for fairness, however, seem to be a purely human trait (Jensen, Call & Tomasello 2007*a*). As Riedl, Jensen, Call & Tomasello (2012) report, vengeance of chimpanzees does not extend to punishing violations of norms without direct consequences to the (potential) punisher. This, however, was the subject of the theoretical model in chapter 4 and the prerequisite to deter free-riding and maintain cooperation in complex human societies. Chapter 4 argues that this specific human trait does have a biological basis, but empirical evidence is still missing. A recent paper by Riedl, Jensen, Call & Tomasello (2015) finds experimental evidence that third-party punishment for

violations of norms is already showing in three-year-old children. The fact that very young children already punish perpetrators might indicate that moral punishment has a biological basis. However, it does not present prove, as socialization in families could already happen in the first three years of life. Research in the spirit of chapter 2 with a sample of children socialized in families with different cultural backgrounds could shed some light on this question.

Bibliography

Aboud, F. E. (1988), *Children and Prejudice (Social Psychology and Society)*, Blackwell Pub.

Aboud, F. E. & Doyle, A. B. (1996), 'Does talk of race foster prejudice or tolerance in children?', *Canadian Journal of Behavioural Science* **28**(3), 161–170.

Ambady, N., Shih, M., Kim, A. & Pittinsky, T. L. (n.d.), 'Stereotype susceptibility in children: Effects of identity activation on quantitative performance', *Psychological Science* **12**(5), 385–390.

Bardsley, N. (2008), 'Dictator game giving: altruism or artefact?', *Experimental Economics* **11**(2), 122–133.

Ben-Ner, A., McCall, B. P., Stephane, M. & Wang, H. (2009), 'Identity and in-group/out-group differentiation in work and giving behaviors: Experimental evidence', *Journal of Economic Behavior & Organization* **72**(1), 153–170.

Benenson, J. F., Pascoe, J. & Radmore, N. (2007), 'Children's altruistic behavior in the dictator game', *Evolution and Human Behavior* **28**(3), 168–175.

Benjamin, D., Choi, J. J. & Strickland, A. J. (2010), 'Social identity and preferences', *American Economic Review* **100**(4), 1913–1928.

Benjamin, D. J., Choi, J. J. & Fisher, G. W. (2010), 'Religious identity and economic behavior', NBER Working Paper no. 15925.

Bernhard, H., Fischbacher, U. & Fehr, E. (2006), 'Parochial altruism in humans', *Nature* **442**(7105), 912–915.

Bettinger, E. & Slonim, R. (2006), 'Using experimental economics to measure the effects of a natural educational experiment on altruism', *Journal of Public Economics* **90**(8-9), 1625–1648.

Binmore, K. & Samuelson, L. (1994), 'An economist's perspective on the evolution of norms', *Journal of Institutional and Theoretical Economics* **150**, 45–63.

Blake, P. R., Piovesan, M., Montinari, N., Warneken, F. & Gino, F. (2014), 'Prosocial norms in the classroom: The role of self-regulation in following norms of giving', *Journal of Economic Behavior & Organization* .

Blake, P. R. & Rand, D. G. (2010), 'Currency value moderates equity preference among young children', *Evolution and Human Behavior* **31**(3), 210–218.

Blanz, M., Mummendey, A., Mielke, R. & Klink, A. (1998), 'Responding to negative social identity: a taxonomy of identity management strategies', *Eur. J. Soc. Psychol.* **28**(5), 697–729.

Bohnet, I. & Frey, B. S. (1999), 'Social distance and other-regarding behavior in dictator games: Comment', *The American Economic Review* **89**(1), 335–339.

Brewer, M. B. (1999), 'The psychology of prejudice: Ingroup love and outgroup hate?', *Journal of Social Issues* **55**(3), 429–444.

Buchan, N. R., Johnson, E. J. & Croson, R. T. A. (2006), 'Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences', *Journal of Economic Behavior & Organization* **60**(3), 373–398.

Burns, J. (2010), Race and trust in a segmented society, *in* M. A. Centeno & K. Newman, eds, 'Discrimination in an Unequal World', Oxford University Press.

Camerer, C. F. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton University Press, Princeton, NJ.

Caswell, J. L., Mallick, S., Richter, D. J., Neubauer, J., Schirmer, C., Gnerre, S. & Reich, D. (2008), 'Analysis of chimpanzee history based on genome sequence alignments', *PLoS Genet* **4**(4), e1000057+.

Chen, Y. & Li, S. X. (2009), 'Group identity and social preferences', *The American Economic Review* **99**(1), 431–457.

Chen, Y., Zhu, L. & Chen, Z. (2013), 'Family income affects children's altruistic behavior in the dictator game', *PLoS ONE* **8**(11), e80419+.

Cherry, T. L., Frykblom, P. & Shogren, J. F. (2002), 'Hardnose the dictator', *The American Economic Review* **92**(4), 1218–1221.

Clark, K. B. & Clark, M. K. (1939), 'The development of consciousness of self and the emergence of racial identification in negro preschool children', *The Journal of Social Psychology* **10**(4), 591–599.

Clark, K. B. & Clark, M. P. (1947), 'Racial identification and preference in negro children.', *Readings in social psychology* **86**, 169–187.

Corenblum, B. & Annis, R. C. (1993), 'Development of racial identity in minority and majority children: An affect discrepancy model.', *Canadian Journal of Behavioural Science* **25**(4), 499–521.

Croson, R. & Gneezy, U. (2009), 'Gender differences in preferences', *Journal of Economic Literature* pp. 448–474.

Dahlman, S., Ljungqvist, P. & Johannesson, M. (2007), 'Reciprocity in young children', SSE/EFI Working Paper Series in Economics and Finance, no. 674, Stockholm School of Economics.

Damon, W. (1977), *The Social World of the Child*, San Francisco: Jossey-Bass.

Damon, W. (1980), 'Patterns of change in children's social reasoning: A two-year longitudinal study', *Child Development* **51**(4), 1010–1017.

Deming, D. & Dynarski, S. (2008), 'The lengthening of childhood', *National Bureau of Economic Research Working Paper Series* pp. 14124+.

Dunbar, R. I. M. (1992), 'Neocortex size as a constraint on group size in primates', *Journal of Human Evolution* **22**(6), 469–493.

Eckel, C. C. & Grossman, P. J. (1998), 'Are women less selfish than men?: Evidence from dictator experiments', *The Economic Journal* **108**(448), 726–735.

Eckel, C. C. & Petrie, R. (2011), 'Face value', *The American Economic Review* pp. 1497–1513.

Engel, C. (2011), 'Dictator games: a meta study', *Experimental Economics* **14**(4), 583–610.

Fehr, E., Bernhard, H. & Rockenbach, B. (2008), 'Egalitarianism in young children', *Nature* **454**(7208), 1079–1083.

Fehr, E. & Gachter, S. (2000), 'Fairness and retaliation: The economics of reciprocity', *The Journal of Economic Perspectives* **14**(3), 159–181.

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

Fong, C. M. & Luttmer, E. F. P. (2009), 'What determines giving to hurricane katrina victims? experimental evidence on racial group loyalty', *American Economic Journal: Applied Economics* pp. 64–87.

Forsythe, R., Horowitz, J. L., Savin, N. E. & Sefton, M. (1994), 'Fairness in simple bargaining experiments', *Games and Economic Behavior* **6**(3), 347–369.

Friesen, J., Arifovic, J., Wright, S. C., Ludwig, A., Giamo, L. & Baray, G. (2012), 'Ethnic identity and discrimination among children', *Journal of Economic Psychology* **33**(6), 1156–1169.

Giamo, L. S. & Wright, S. C. (2008), 'East asian children's intergroup contact experiences: An investigation of outgroup attitudes, sharing behaviour and anxiety as a mediator', Unpublished MA thesis. Simon Fraser University, Burnaby, BC.

Goeree, J. K., McConnell, M. A., Mitchell, T., Tromp, T. & Yariv, L. (2010), 'The 1/d law of giving', *American Economic Journal: Microeconomics* pp. 183–203.

Goette, L. F., Huffman, D. & Meier, S. (2006), 'The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups', *Social Science Research Network Working Paper Series* .

Guala, F. & Mittone, L. (2010), 'Paradigmatic experiments: The dictator game', *The Journal of Socio-Economics* **39**(5), 578–584.

Gummerum, M., Hanoch, Y. & Keller, M. (2008), 'When child development meets economic game theory: An interdisciplinary approach to investigating social development', *Human Development* **51**(4), 235–261.

Gummerum, M., Keller, M., Takezawa, M. & Mata, J. (2008), 'To give or not to give: Children's and adolescents' sharing and moral negotiations in economic decision situations', *Child Development* **79**(3), 562–576.

Gummerum, M., Takezawa, M. & Keller, M. (2009), 'The influence of social category and reciprocity on adults' and children's altruistic behavior', *Evolutionary Psychology* **7**(2), 295–316.

Güth, W., Schmittberger, R. & Schwarze, B. (1982), 'An experimental analysis of ultimatum bargaining', *Journal of Economic Behavior & Organization* **3**(4), 367–388.

Häger, K. (2010), 'Envy and altruism in children', Jena Economic Research Papers, no. 2010,063.

Häger, K., Oud, B. & Schunk, D. (2012), 'Egalitarian envy: Cross-cultural variation in the development of envy in children', Jena Economic Research Papers, no. 2012 - 059.

Harbaugh, W. T. & Krause, K. (2000), 'Children's altruism in public good and dictator experiments', *Economic Inquiry* **38**(1), 95–109.

Harbaugh, W. T., Krause, K. & Liday, S. G. (2002), 'Bargaining by children', Working paper 2002-04, University of Oregon, Department of Economics.

Harbaugh, W. T., Krause, K. & Vesterlund, L. (2007), 'Learning to bargain', *Journal of Economic Psychology* **28**(1), 127–142.

Harrod, W. J. (1983), 'In-Group bias in the minimal organizational setting', *Simulation & Games* **14**(3).

Hastings, P. D., Utendale, W. T. & Sullivan, C. (2007), The socialization of prosocial development, *in* J. E. Grusec & P. D. Hastings, eds, 'Handbook of Socialization: Theory and Research', Guilford Press, chapter 25, pp. 638–664.

Hertwig, R. & Ortmann, A. (2001), 'Experimental practices in economics: A challange for psychologists?', *Behavioral and Brain Sciences* **24**(3).

Hoffman, E., Mccabe, K. & Smith, V. L. (1996), 'Social distance and Other-Regarding behavior in dictator games', *The American Economic Review* **86**(3), 653–660.

Hofstede, G. (1980), *Culture's Consequences*, University of California Press, Berkley, CA.

Hornsey, M. J. & Hogg, M. A. (2000), 'Assimilation and diversity: An integrative model of subgroup relations', *Personality and Social Psychology Review* **4**(2), 143–156.

Hornsey, M. J. & Hogg, M. A. (2002), 'The effects of status on subgroup relations', *British Journal of Social Psychology* **41**(2), 203–218.

Houle, R. (2011), 'Recent evolution of immigrant-language transmission in Canada', *Statistics Canada Catalogue no. 11-008-X - Canadian Social Trends* .

House, B. R., Henrich, J., Brosnan, S. F. & Silk, J. B. (2012), 'The ontogeny of human prosociality: behavioral experiments with children aged 3 to 8', *Evolution and Human Behavior* **33**(4), 291–308.

Houser, D. & Schunk, D. (2009), 'Social environments with competitive pressure: Gender effects in the decisions of german schoolchildren', *Journal of Economic Psychology* **30**(4), 634–641.

Jensen, K., Call, J. & Tomasello, M. (2007*a*), 'Chimpanzees are rational maximizers in an ultimatum game', *Science* **318**(5847), 107–109.

Jensen, K., Call, J. & Tomasello, M. (2007*b*), 'Chimpanzees are vengeful but not spiteful', *Proceedings of the National Academy of Sciences* **104**(32), 13046–13050.

Kahneman, D., Knetsch, J. L. & Thaler, R. (1986), 'Fairness as a constraint on profit seeking: Entitlements in the market', *The American Economic Review* **76**(4), 728–741.

Katz, D. & Braly, K. (1933), 'Racial stereotypes of one hundred college students', *Journal of abnormal and social psychology* **28**(3), 280–290.

Kench, B. T. & Niman, N. (2009), 'Of altruists & thieves', *Social Science Research Network Working Paper Series* .

Kluczniok, K. (2012), *Die vorzeitige Einschulung*, Waxmann Verlag.

Kollock, P. (1998), Transforming social dilemmas: group identity and co-operation, *in* P. A. Danielson, ed., 'Modeling Rationality, Morality, and Evolution', Oxford University Press, New York, pp. 185–209.

Kuzmics, C. & Rodriguez-Sickert, C. (2007), The evolution of moral codes of behavior. **URL:** *http://ssrn.com/paper=910292*

Lincove, J. A. & Painter, G. (2006), 'Does the age that children start kindergarten matter? evidence of Long-Term educational and social outcomes', *Educational Evaluation and Policy Analysis* **28**(2), 153–179.

List, J. A. (2007), 'On the interpretation of giving in dictator games', *Journal of Political Economy* **115**(3), 482–493.

Long, S. J. & Freese, J. (2005), *Regression Models for Categorical Dependent Variables Using Stata, Second Edition*, 2 edn, Stata Press.

Malti, T., Gummerum, M., Ongley, S., Chaparro, M., Nola, M. & Bae, N. Y. (2016), ' who is worthy of my generosity? recipient characteristics and the development of children's sharing', *International Journal of Behavioral Development* **40**(1), 31–40.

Martin, C. L., Ruble, D. N. & Szkrybalo, J. (2002), 'Cognitive theories of early gender development', *Psychological bulletin* **128**(6), 903–933.

McGillicuddy-de Lisi, A. V., Watkins, C. & Vinchur, A. J. (1994), 'The effect of relationship on children's distributive justice reasoning', *Child Development* **65**(6), 1694–1700.

Messick, D. M. (1993), Equality as a decision heuristic, *in* B. A. Mellers & J. Baron, eds, 'Psychological perspectives on justice: Theory and applications', Cambridge University Press, Cambridge, pp. 11–31.

Moore, C. (2009), 'Fairness in children's resource allocation depends on the recipient', *Psychological Science* **20**(8), 944–948.

Murnighan, K. J. & Saxon, M. S. (1998), 'Ultimatum bargaining by children and adults', *Journal of Economic Psychology* **19**(4), 415–445.

Nesdale, D., Maass, A., Griffiths, J. & Durkin, K. (2003), 'Effects of in-group and out-group ethnicity on children's attitudes towards members of the in-group and out-group', *British Journal of Developmental Psychology* **21**(2), 177–192.

Olson, K. R. & Spelke, E. S. (2008), 'Foundations of cooperation in young children', *Cognition* **108**(1), 222–231.

Paulus, M. (2016), 'Friendship trumps neediness: The impact of social relations and others' wealth on preschool children's sharing', *Journal of Experimental Child Psychology* **146**, 106–120.

Peterson, D. & Wrangham, R. (1997), *Demonic Males: Apes and the Origins of Human Violence*, 1 edn, Mariner Books.

Rao, N. & Stewart, S. M. (1999), 'Cultural influences on sharer and recipient behavior', *Journal of Cross-Cultural Psychology* **30**(2), 219–241.

Riedl, K., Jensen, K., Call, J. & Tomasello, M. (2012), 'No third-party punishment in chimpanzees', *Proceedings of the National Academy of Sciences* **109**(37), 14824–14829.

Riedl, K., Jensen, K., Call, J. & Tomasello, M. (2015), 'Restorative justice in children', *Current Biology* **25**(13), 1731–1735.

Robson, A. J. (2002), 'Evolution and human nature', *The Journal of Economic Perspectives* pp. 89–106.

Rochat, P. (2011), 'Possession and morality in early development', *New Directions for Child and Adolescent Development* **2011**(132), 23–38.

Rochat, P., Dias, M. D. G., Liping, G., Broesch, T., Passos-Ferreira, C., Winning, A. & Berg, B. (2009), 'Fairness in distributive justice by 3- and 5-Year-olds across seven cultures', *Journal of Cross-Cultural Psychology* **40**(3), 416–442.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. (1991), 'Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study', *The American Economic Review* **81**(5), 1068–1095.

Samuelson, L. (1997), *Evolutionary games and equilibrium selection*, MIT Press series on economic learning and social evolution, MIT Press.

Schelling, T. C. (1960), *The strategy of conflict.*, Harvard University Press, Cambridge, MA.

Shih, M., Pittinsky, T. L. & Ambady, N. (1999), 'Stereotype susceptibility: Identity salience and shifts in quantitative performance', *Psychological Science* **10**(1), 80–83.

Smith, V. L. (2007), *Rationality in Economics: Constructivist and Ecological Forms*, 1 edn, Cambridge University Press.

Spielman, D. A. (2000), 'Young children, minimal groups, and dichotomous categorization', *Personality and Social Psychology Bulletin* **26**(11), 1433–1441.

Stanford, C. B. (2001), *The Hunting Apes: Meat Eating and the Origins of Human Behavior*, Princeton University Press.

Tajfel, H. & Turner, J. (1986), The social identity theory of intergroup behavior, *in* S. Worchel & W. Austin, eds, 'Psychology of intergroup relations', Nelson-Hall, Chicago.

Triandis, H. C. (1995), *Individualism And Collectivism (New Directions in Social Psychology)*, Westview Press.

Tversky, A. & Kahneman, D. (1981), 'The framing of decisions and the psychology of choice', *Science* **211**(4481), 453–458.

Wright, S. C. & Taylor, D. M. (1995), 'Identity and the language of the classroom: Investigating the impact of heritage versus second language instruction on personal and collective self-esteem.', *Journal of Educational Psychology* **87**(2), 241–252.

Appendix

Additional Material for Chapter 2

Scripts

A total of 13 different scripts were used, with different treatments and different orders of experiments.

Script 01 Sorting task first. Simultaneous sharing with trial three under increased anonymity, no profile cards used

Script 02 Sorting task first. Simultaneous sharing with trial two under increased anonymity, no profile cards used

Script 03 Sorting task first. Simultaneous sharing with trial three under increased anonymity, profile cards used with additional information on hypothetical others

Script 04 Sorting task first. Simultaneous sharing with trial two under increased anonymity, profile cards used with additional information on hypothetical others

Script 05 Dictator game first. Simultaneous sharing with trial three under increased anonymity, no profile cards used

Script 06 Dictator game first. Simultaneous sharing with trial two under increased anonymity, no profile cards used

Script 07 Dictator game first. Simultaneous sharing with trial three under increased anonymity, profile cards used with additional information on hypothetical others

Script 08 Dictator game first. Simultaneous sharing with trial two under increased anonymity, profile cards used with additional information on hypothetical others

Script 09 Sorting task first. Sequential sharing, no profile cards used

Script 10 Sorting task first. Sequential sharing, profile cards used with additional information on hypothetical others

Script 11 Dictator game first. Sequential sharing, no profile cards used

Script 12 Dictator game first. Sequential sharing, profile cards used with additional information on hypothetical others

Script 20 Dictator game first. Sequential sharing, no profile cards used

Script 20 was used only at the end of the field work to test another experiment (a reciprocity game) instead of the sorting task. The dictator game was unaffected - it was run in the same way as in script 09.

The following pages document script 1 as an example.

ICAB Tester Script

There are multiple scripts – all have a specific order of tasks to be performed. Make sure you take the script corresponding to the answer sheet for the next kid.

Protocol design 2007-08 for
- sorting task: Steve Wright and Lisa Giamo, Department of Psychology
- sharing task: Andreas Ludwig, Department of Economics
Simon Fraser University, Burnaby, BC

Materials: log sheet, photographs, result sheets for recording data, digital camera, printer, a box, stickers, profile cards and envelopes

Study Set-Up

1. Before beginning, make sure child has appropriate informed consent forms signed by a parent/guardian.
2. Write initials and time next to appropriate participant on log sheet.
3. Fill out **ALL** information at the top of participant information and result sheet: participant and class number (make sure this matches the number on the log sheet), your initials; date, time; age, gender and ethnicity of the kid
4. Make sure all photographs are present and in good condition.
5. Make sure digital camera is in working order.
6. Make sure the printer is on and working.

Procedure

As participant arrives, introduce yourself and ask child's name.

> Hi, my name is _____, and I will be asking you some questions today.
>
> What's your name?

Wait for response.

> And how old are you?

Wait for response. Feel free to say things like 'Very cool' or anything similar to let child know you are interested and engaged.

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

Now you are going to take a photograph of the child.

> _Child's name_, before we start, would it be alright if I take a picture of you? We are going to be using it in some of the games we play, and then give it to you to take home with you.

Wait for response.

> Awesome.

Take picture.

> That was an excellent picture!

Print the picture.

> OK, _child's name_ let me tell you a little about what we are going to be doing today. We are going to play a few different games. All of them use pictures of other people and some will use your picture too.

I. Sorting task

Use the package of 12 photos of individual kids.

!!! IMPORTANT -- BEFORE YOU START !!!

- Put the kid's picture into the pile of photos.
- Use exactly the same procedures for each sorting trial.
- Shuffle the photos before each sorting trial.
- Begin each trial by laying out all photographs (including the child's own) facing the child.
- If child is having difficulty with a trial, help them out with the prepared prompts.
- If the child still does not respond after the prompts, move on to the next item. **DO NOT INFLUENCE THEIR RESPONSES IN ANY OTHER WAY**.
- You may need to remind the child to pay attention and/or slow them down if they seem not to be listening to the specific question or not really looking over all the pictures.
- Say only "thank you" after each response. But say "great, well done!" after you finish each task.

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

For the first game, I'm going to put your picture into this pile of pictures of other children. These are kids from another school, so you won't know any of them. But they should look sort of like some of the kids at your school. I'm going to ask you to look at all these kids, and I will ask you some questions about them. Remember, if you are confused or you don't know exactly what you are supposed to do, just tell me or ask a question, OK?

Wait for response.

OK. Are you ready to begin?

Wait for response.

Great!!

Shuffle photographs and lay them out in front of the child.

Can you reach all of the pictures?

If no, re-organize them so that s/he can.

Practice trials:
These should convince you that the child understands the tasks

Ok, _child's name_ look at all these kids, and point to all the ones who are **girls** and leave all the ones who are **not girls** on the table.

Thank you!

Now look at all these kids, and point out all the ones who are **boys** and leave all the ones who are **not boys** on the table.

Thank you!

Great, well done!

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

Actual test trials:
For every item, record the child's choices and any comments you may have on the result sheet

OK, child's name I know that you don't know any of these kids, but I want you to imagine that they are kids in your class.

Look at all the kids, and point to all the kids that you think **are nice to other kids** and leave all the kids who **are not nice to other kids** on the table. You can select some of the kids, all or none of them. If you want to pick all or none of them, you may say "all" or "none".

Thank you!

Are **smart**… are **not smart**

Are **happy** … are **not happy**

Like to go to **school**… **do not like** to go to **school**

Have **lots of friends**… do **not** have **lots of friends**

Are **bad**… are **not bad**

Always need help from other kids and the teachers… do **not need help** from others

Work hard at school… do **not work hard** at school

Are **mean** to other kids… are **not mean** to other kids

Are **helpful**… are **not helpful**

Can **read well**… **cannot read**

Optional Optional

Prompts

If the child seems bored at any time, you can say things like

There are only a few more to do.

We're almost done!

We'll be doing something a bit different next.

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

SORTING TASK: Feelings

!!! IMPORTANT -- BEFORE YOU START !!!

Remove the child's picture from the pile of photos and put it right in front of the child.

> For the next part, we are going to be doing something a little different. We are going to take your picture out of the pile of other pictures and put it over here (place picture in front of the child).

Shuffle remaining 12 pictures.

> OK, _child's name, this time, I am going to ask you some questions about how these other children would make you feel. Look at all these kids again and imagine that you are playing with them. Pick out all the kids who would **make you feel safe** and put them **on top of your picture close to you**, and leave all the ones who would **not make you feel safe** on the table.

Shuffle pictures.

> This time, I want you to look at all these kids again pick out all the kids who would **make you feel scared** and put them here (*point at table in front of you)*, **far away from you**, and leave all the ones who would **not make you feel scared** on the table

Shuffle pictures.

> And finally, I want you to look at all these kids again pick out all the kids who would **make you feel worried** and put them here (*point at table in front of you)*, **far away from you**, and leave all the ones who would **not make you feel worried** on the table

SORTING TASK: Similarity and friendship

- Follow an analogous procedure as in the other sorting tasks.
- Reshuffle pictures.

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

!!! New intro !!!

I want you to look at all these kids again and tell me which of these kids you **think are a lot like you**, and put them on top of your picture, and leave the kids that are **not like you** on the table.

Now I want you to look at all these kids one last time. I want you to pick all the kids that you **would like to have as a good friend** and put them on top of your picture, and leave all the kids who you **do not want as a good friend** on the table.

II. Sharing Task

!!! IMPORTANT -- BEFORE YOU START !!!

- Make sure you have the 3 envelopes with stickers and 3 sets of picture envelopes (pictures of boys for testing boys and girls for girls)
- Introduce the sharing task in the following way **(VERBATIM!)**

OK, child's name. We still have a real fun thing for you to do. We won't need your picture anymore so we will put it over here for now.

- Take out of the first set of stickers
- Spread the stickers on the desk so they can easily be counted, but close to the kid (suggesting the stickers belong to him/her).
- **Do not count** the stickers; just casually spread them so that each individual sticker can be seen.

Child's name , here you get 12 stickers. They are yours to keep. OK?

Wait for response.

Take first 3 envelopes labeled "**1 …**" on the back, shuffle them.

But before you take them home, there are 3 other boys (girls) who don't have any stickers and would like to have some as well. Here are pictures of kids that look like the 3 other boys (girls).

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

Place the 3 envelopes on the table, photos facing the child.

> Please decide how many stickers you want to share with each of those 3 kids and how many you want to keep for yourself. It is completely up to you how you split up the stickers.

Pause.

> You can keep them all if you like, or you can give stickers to one of these boys (girls).
>
> Or you can give some stickers to two of these boys. Or you can give some stickers to all of them.
>
> If you share stickers with any of these boys (girls), we will put them into an envelope with their picture and I will give each envelope to a kid like the one on the picture afterwards.
>
> What would you like to do?

Wait for the kid to indicate if she/he wants to share.

If he/she indicates he/she wants to share say,

> Then put the stickers that you want to share with each boy (girl) on top of his (her) picture. Go ahead and take your time to share the stickers as you want. If you are finished, please say "Finished!". I will wait for you to tell me, OK?

Let her/him finish distributing the stickers. **Do not prompt "are you finished"** but wait for the kid to indicate that s/he is done.

> Thanks. Let's put the stickers into the envelopes now. Please put the ones that you want to keep for yourself into your envelope.

Give empty envelope that contained the first set of stickers to the child.

!!! Important !!!

- Put the stickers into the correct envelopes.
- Put the target kids' envelopes into the box.
- Make sure to put the stickers the kid didn't want to share into the blank envelope.

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

- Do not leave any stickers on the table.

Thank you!

Pick up the next 3 envelopes (with the right number on the back), and shuffle.

Take out the next set of stickers.

Now, I have another pack of stickers and we are going to do the same thing again.

It is completely up to you how you share your stickers.

As you talk, spread the stickers and put the envelopes down, photos facing the child.

What would you like to do?

Wait for the kid to indicate if she/he wants to share.

If he/she indicates he/she wants to share say,

Then put the stickers that you want to share with each boy (girl) on top of his (her) picture. Go ahead and take your time to share the stickers as you want. If you are finished, please say "Finished!". I will wait for you to tell me, OK?

Let her/him finish distributing the stickers. **Do not prompt "are you finished"** but wait for the kid to indicate that s/he is done.

Thanks. Let's put the stickers into the envelopes now. Please put the ones that you want to keep for yourself into your envelope.

Give empty envelope that contained the first set of stickers to the child.

!!! Important !!!

- Put the stickers into the correct envelopes.
- Put the target kids' envelopes into the box.
- Make sure to put the stickers the kid didn't want to share into the blank envelope.
- Do not leave any stickers on the table.

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

> Thank you!

Pick up the next 3 envelopes (with the right number on the back), and shuffle.

Take out the next set of stickers.

> Now, I have another pack of stickers and we are going to do the same thing again.
>
> It is completely up to you how you share your stickers.

As you talk, spread the stickers and put the envelopes down, photos facing the child.

> Well, I think this time you know how it works. If you want to share any of your stickers, just put them into the right envelope and put the envelopes into the box with all the others. And put the stickers you want to keep into your envelope.
>
> I will not look at what you are doing, so nobody will ever know how you share the stickers, not your teacher, not your friends, not your parents, not even me. OK?

Wait for response. Stop the kid if he (she) starts to share before you finished the instructions and turned around.

> Go ahead and take your time to share the stickers as you want. If you are finished, please say "Finished!" I will wait for you to tell me, OK?

Turn around and wait for the kid to distribute the stickers and to indicate that the task is completed. Prompt with prompts given below if needed. If the kid indicates to be finished, prompt **before** turning around:

> Did you put all the stickers away? I don't want to see any stickers on the table anymore! And please make sure that the envelope for the other boys (girls) are in the box.

!!! Important !!!

- Make sure the kid clearly indicates that all stickers are in the respective envelopes and the sharing envelopes are in the box **before you turn around**.

Thank you!

Prompts

Optional

Again, if the child is having difficulty with the task, encourage the kid with one of the following prompts.

Optional

It is completely up to you how you share the stickers.

You can keep them all if you like, or you can give some of them to one of these boys (girls), or you can give some to two of these boys (girls), or you can give some to all of these boys (girls).

!!! Important !!!

Do not influence their responses in any other way. **Do not ask "are you finished" but wait for the child to indicate it.**

Wrap-up

End of study.

OK __child's name_____, we are all done!!! You did a really great job today. Thanks for all of your help. Don't forget to take your picture and your stickers with you to take home.

Make sure all the stickers that were not shared and the picture are in the child's envelope, seal the envelope, write the child's name onto it and give the envelope to the child.

I will take you back to your classroom now. Thanks again!!!

Clean-up

- Make sure to record any inconsistencies on result sheet. be sure the result sheet has participant number etc. Put it into the folder.

Script 1: Sorting -> Sharing, NO profiles, sharing: open-open-anonymous
Corresponds to result sheets **1**, **1**(2), **1**(3)…

- Delete the child's picture from the camera.
- Take the envelopes out of the box and count the number of stickers shared in each round
- Look at the next results sheet and select the appropriate script and set of envelopes for the next trial (boys or girls).
- Close the alternative set of envelopes (opposite gender), shuffle them and put them into the box

END OF THE EXPERIMENT

Additional Material for Chapter 3

Coding of Neighbourhood Characteristics

Neighbourhood characteristics are based on public-use aggregates of the Census of Population "long form," administered by Statistics Canada to one in five households in 1996 and 2001. The lowest level of geography for which Statistics Canada produced aggregate statistics based on the 2001 Census is a Dissemination Area (DA). DAs are geographic areas designated for the collection of Census data. DAs are composed of one or more neighbouring blocks with a population of 400 to 700 persons.

We link postal codes to DAs using Statistics Canada's Postal Code Conversion File (PCCF). The PCCF contains a complete longitudinal correspondence between postal codes and DAs (postal codes are occasionally retired and subsequently recycled). Postal codes are smaller than DAs and usually lie entirely within a DA. In cases where postal code boundaries span multiple DAs, we use the PCCF's Single Link Indicator (which identifies the best link to an DA) to link to a unique DA.

We were unable to assign DA-level characteristics to residential postal codes in 20 cases. This arose when residential postal codes did not appear in the PCCF (most likely due to misreported postal codes), or when DA-level characteristics were suppressed by Statistics Canada for confidentiality reasons.