Biases and Discrimination: An Economic Analysis using Lab and Field Experiments

Submitted by Graeme Pearce, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Economics, in September 2016.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other university.

Signature:

Abstract

This thesis uses laboratory and field experiments to examine the underlying motivations that drive biased and discriminatory behaviour. Its focus is on the differential treatment of others that stems from individuals' preferences for particular social and ethnic groups. The unifying theme of this thesis is the exploration of how such discriminatory tastes can manifest themselves within individuals' social and other–regarding preferences, determining the extent to which they care about the welfare of others. The prevalence and implications of these types of preferences are considered in both market and non–market settings.

This thesis is dedicated to my daughter, Isabella.

Acknowledgements

Whilst writing this thesis I have benefited from very many helpful discussions and suggestions from my supervisor, Brit Grosskopf. A special thanks goes to her, as her guidance and support has contributed immeasurably to my personal development and to the progress of this thesis.

I also gratefully acknowledge the 1+3 ESRC studentship granted to me by the University of Birmingham, and the financial support provided by the University of Exeter Business School. Without this financial support I would not have been able to complete, or even begin, this project.

Finally, I would like to acknowledge the contributions of my wife and best friend, Debra. Without Debra's encouragement I would never have undertaken this project, and without her unconditional support it is unlikely that I would have succeeded in completing it.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis uses both laboratory and field experiments in order to advance our understanding of the motivations that underpin biased and discriminatory behaviour. Its primary focus is the analysis of 'taste–based' discrimination, or the differential treatment of others stemming from an individual's preference for a particular social or ethnic group (Becker, 1971). The unifying theme of this thesis is the exploration of how such discriminatory tastes can manifest themselves within individuals' social and other–regarding preferences, affecting the extent to which they care about the welfare of others (Chen & Li, 2009).

This short introductory chapter is designed to briefly set the intellectual landscape, and is organised as follows. Section 1.1 motivates the experimental methods employed in this thesis, and summarises the recent debate regarding the strengths and weaknesses of lab and field methodologies. Section 1.2 briefly discusses common laboratory and field experimental methods used in previous studies of discrimination. Section 1.3 gives an overview of Chapter 2. Section 1.4 outlines Chapter 3 and Chapter 4 is introduced in Section 1.5.

1.1 Motivating the experimental method

Experiments are particularly well suited to the task of examining taste–based discrimination and testing the behavioural theories regarding how these tastes might be expressed. Experiments afford the researcher the control to identify causal effects and allow the consideration of empirical regularities that might support, or refute, theoretical predictions. Importantly, experiments allow behaviour to be examined across institutions (Smith, 1994) in a tightly controlled setting. The laboratory provides an ideal environment in which theory can be tested rigorously, granting an abstract environment in which to reduce the possible confounds that are found in field data. This grants the researcher a level of control unattainable in observational data.

However, many researchers have begun to question the robustness of results obtained from the laboratory, and have begun to recognise their limitations. As highlighted by Levitt & List (2007), the laboratory setting has many inherent properties which may have undesirable consequences for behaviour, or that may affect the generalisability of any conclusions. Features such as experimenter scrutiny, the self–selection of participants into the lab and the artificial restriction placed on choice sets may all influence the way in which subjects behave. Studies have now begun to explore the implications of these attributes in more detail (Winking & Mizer, 2013; Stoop *et al.*, 2012; Stoop, 2014). Section 4.2 of Chapter 4 expands on this discussion considerably.

The criticisms and concerns raised about the generalisability and interpretability of laboratory experiments has triggered a heated debate in the literature (Camerer, 2015; Al-Ubaydli & List, 2015). This has led many researchers to now regard natural field experiments, those that are conducted in natural settings where subjects are unaware their behaviour is being scrutinised, as the gold standard for testing economic theory and establishing causal treatment effects (Al-Ubaydli & List, 2015).[1] Although some researchers view lab and field methodologies as competing, the majority view them as complementary.[2] The 'scientific view' as defined by Camerer (2015), and the view taken in this thesis, is that all empirical work, experimental or otherwise, contributes to a general economic theory that seeks to explain how individuals respond to incentives.

1.2 Experiments in the discrimination literature

Experiments have been used to study discrimination for decades, with the literature's main objective largely being to distinguish taste–based discrimination, as proposed by Becker (1971), from statistical discrimination, as formulated by Phelps (1972). The ability to distinguish between these competing explanations is important for determining market efficiency (Schwab, 1986) and the extent to which individuals are behaving

[1] Harrison & List (2004) provide a comprehensive discussion of the various types of field experiment.

[2] Al Roth provides a well rounded discussion of these issues in his slides from the 2006 ASSA meeting, which can be found at the following address: https://web.stanford.edu/ alroth/papers/LabVersusField.Roundtable.ASSA.Jan.2006.pdf.

rationally. This is therefore important for the formulation of policy proposals.

Previous work has employed various laboratory and field experiment methodologies, including correspondence and audit methods, but has also utilised observational data to shed light on these two very different motives. The literature has considered discrimination against a number of identities in a range of settings, including gender and racial discrimination in labour markets (Goldin & Rouse, 2000; Bertrand & Mullainathan, 2004), ethnic discrimination in housing markets (Ahmed & Hammarstedt, 2008) and discrimination against the disabled in the market for car repairs (Gneezy *et al.*, 2012).

In the lab, there is a considerable body of research that focuses on discriminatory behaviour. Researchers have employed methods such as the exogenous revelation of surnames (Fershtman & Gneezy, 2001) and the country of origin of subjects (Whitt & Wilson, 2007), in order to determine if subjects condition their behaviour on the identity of others. However, to better understand taste–based discrimination, experimenters have turned to study artificially induced identities, categorising subjects into random, or arbitrarily assigned, social groups and then studying how they interact. These methods are often seen as preferable to the study of natural identities as they allow the experimenter to control the identity that guides behaviour, and thus avoid the complexities and confounds inherent in the study of natural identity. It is argued that such methods allow the study of identity effects *per se* (Chen & Chen, 2011). This work, building on the social psychology work of Tajfel *et al.* (1971) and Tajfel & Turner (1986), was pioneered in economics by Charness *et al.* (2007) and Chen & Li (2009). However, this literature has faced strong criticisms from Guala & Filippin (2015), Zizzo (2010) and Zizzo (2012), who suggest the results of laboratory studies of discrimination are often indistinguishable from experimenter demand effects.

In the field, the correspondence method represents one of the most well utilised methodologies in the literature. Researchers using this method produce large numbers of identical CVs or letters and exogenously vary either the ethnicity, nationality or gender of the applicant or author through the use of names, addresses or photos. This method was brought to prominence by Bertrand & Mullainathan (2004), who employed it to study discrimination against black job applicants. However, it has also been used to study discrimination against those viewed as being less attractive (Ruffle & Shtudiner, 2015), foreign nationals (Booth *et al.*, 2012) and Muslims (Ahmed & Hammarstedt, 2008). Researchers have also used 'lost letter' experiments, which fall into this category of experiment, with similar successes (Ahmed, 2010*a*). A problem with this type of study, as is often highlighted, is the inability to determine if treatment differentials are a result of taste–based or statistical discrimination, or even a consequence of discrimination at all.

The audit study method also represents a workhorse within the discrimination literature. This method employs pairs, or groups, of individuals to complete simple standardised interactions, such as bargaining for a car or interviewing for a job.[3] The experimenter varies the characteristic of interest, such as gender or ethnicity, within the pair or group of employed individuals. This method has been widely implemented, particularly in the study of hiring decisions, but more recent work has also focused on the treatment of homosexuals and the disabled (Gneezy *et al.*, 2012). However, many assumptions about the unobservables that might be correlated with the variable of interest must be made in order to distinguish behaviour from statistical discrimination (Heckman, 1998).

1.3 Overview of Chapter 2

Chapter 2 builds on the social psychology research of Tajfel *et al.* (1971) and Tajfel & Turner (1986). Its objective is to consider how a common sense of identity, or sense of self, impacts reciprocal behaviour and in doing so examines a potential behavioural channel through which discrimination might be expressed.

The experiment induces an artificial sense of identity into subjects by utilising a group problem solving task. Following Chen & Li (2009) subjects are randomly assigned to one of two groups, and in an initial stage can chat to their group members in order to identify abstract paintings. Once this is complete, subjects play the gift–exchange game of Fehr *et al.* (2007), playing as either *Principals* or *Agents*. *Principals* offer *Agents* wages and ask for effort levels in an 'incomplete contract' and are given the opportunity to reciprocate *Agents*' choices of effort by paying unenforceable bonuses. Using a within–subjects design the experiment varies whether subjects play with an in–group player, or an out–group player.

The results suggest that an induced sense of identity has no effect on behaviour, and subjects do not differentiate their behaviour conditional on the identity of others. We conclude that the identity inducement procedure may have interacted with the framing of the experiment, as although subjects playing as *Agents* do regard their social group as being important, those playing as *Principals* do not.

[3]See Riach & Rich (2002) for a survey of audit studies.

This chapter also provides a conceptual replication of Fehr *et al.* (2007), and includes a treatment without identity. Interestingly, we find that reciprocal behaviour in our replication is significantly smaller than in the original study. We use the original data of Fehr *et al.* (2007) to structurally estimate inequity aversion parameters and find subjects in the replication to be around 20% less inequity averse than in the original.

1.4 Overview of Chapter 3

Chapter 3 presents a field experiment that examines the nature and extent of discrimination faced by Muslims. This is achieved by studying the behaviour of a previously unstudied demographic: the poorest people in England, the residents of two housing estates in the town of Rochdale. This chapter is interested in examining why individuals from this particular demographic are likely to vote for and support anti–Muslim policies.

In the experiment, subjects first divide £10 between two strangers. They then play a Dictator Game, dividing £10 between themselves and a different stranger, with stake sizes equal to up to 18% of their weekly income. The ethnicity of the receivers in each game is subtly varied by providing subjects with surnames randomly drawn from the electoral register for the surrounding area. The experiment includes a treatment where receivers' surnames are withheld, providing a baseline from which behaviour can be parsed into either in–group favouritism or out–group negativity. This is an important behavioural distinction, which the discrimination literature typically overlooks.

The results suggest that this demographic may support policies that discriminate against Muslims as a result of out–group negativity, rather than a consequence of in–group favouritism. We advance the literature on discrimination through the estimation of a structural model of group–contingent social preferences (Chen & Li, 2009), which we exploit to perform counterfactual simulations. The results provide a potential explanation for the negative attitudes towards Muslims and a possible motivation behind the documented discriminatory behaviour against them.

1.5 Overview of Chapter 4

The purpose of Chapter 4 is to examine the prevalence, and extent, of other–regarding preferences in transactions completed in a highly competitive market place, and determine the role played by reputational concerns in fostering other–regarding behaviour. It also examines how ethnicities interact and influence pro–social behaviour. This is done using a natural field experiment.

To achieve this, we employed testers of varying ethnicity to take a number of predetermined taxi journeys. In each case we endowed them with only 80% of the expected fare. Testers revealed the amount they could afford to pay to the driver mid-journey and asked for a portion of the journey for free. In a 2×2 between–subjects design we vary the length of the journey and whether drivers have reputational concerns or not.

We find that the majority of drivers give at least part of the journey for free and over 25% complete the journey. Giving is found to be proportional to the length of the journey, and the drivers' reputational concerns do not explain their behaviour. Evidence of strong out–group negativity against black testers by both white and South–Asian drivers is also reported. In order to link our empirical analysis to behavioural theory we structurally estimate the parameters of a number of utility functions. We find that drivers' other–regarding preferences are group–contingent, as proposed by Chen & Li (2009).

This chapter makes a number of contributions. First, it adds to the debate surrounding the generalisability of results obtained from laboratory experiments. Second, it contributes to the other–regarding preference literature, by providing the first field evidence that such preferences can, and do, influence behaviour in market settings (DellaVigna, 2009). Finally, it contributes to the discrimination literature, and sheds further light on how discrimination can manifest itself within our preferences.

Chapter 2

Group Identity and Reciprocity: A Laboratory Experiment

2.1 Introduction

Since the work of Akerlof (1982), reciprocity has been a central topic of research for many economists. The notion of 'gift-exchange' between economic agents, that employers pay higher wages to induce greater efforts from workers, has become a stylised fact, with a number of authors suggesting that reciprocal motives are prevalent. Researchers have suggested that reciprocity can stop markets clearing (Fehr *et al.*, 1993, 1998*a*), explain why individuals reject bargaining offers that would otherwise make them better off (Fehr & Schmidt, 1999), and provide an explanation for why people contribute to public goods and punish those who don't (Fehr & Gächter, 2000*a*). There is also field evidence to support the gift-exchange hypothesis (Falk, 2007; Kube *et al.*, 2012), although the field experiments of List (2006), and Gneezy & List (2006) question the generalisability of the laboratory results.

Reciprocity has been studied in many different settings (Berg *et al.*, 1995; Fehr *et al.*, 1998*b*) but has gained particular traction in the incomplete contracting literature, notably in its use as a contract enforcement device (Fehr & Gächter, 2000*b*): workers in experimental labour markets appear to exert greater levels of 'effort' when experimental employers are able to reward or punish their behaviour. Andreoni *et al.* (2003) suggest that punishments, rather than rewards, provide the most powerful incentives, although the more recent work of Fehr *et al.* (2007) highlights how discretionary, reciprocal rewards can be just as effective. The field experiments of Kube *et al.* (2012) support this result. In support of the conclusions of Fehr *et al.* (1997), it appears as though strong, positive reciprocity can act as a powerful norm enforcement device and result in significant efficiency gains. Even if only a fraction of the population exhibit reciprocal preferences, theory suggests that even purely selfish agents can be induced to exert more effort because of the expectation of reciprocity from their employer (Fehr & Schmidt, 1999).

Fehr *et al.* (2007) (herein FKS) is one of the most well published, and well cited studies within the incomplete contracting literature.[1] FKS is a study designed to compare the performance of 'complete' incentive contracts with 'incomplete' bonus contracts experimentally. Complete contracts are enforced normatively, whilst incomplete contracts are enforced by reciprocity. Subjects are first given a role as either a *Principal* or an *Agent*: *Principals* select which contract to offer to an *Agent*, complete or incomplete, and then specify the contract details. By offering an incentive contract, the *Principal* offers the *Agent* a wage, and a potential punishment if they choose too low an effort level, which is verified with some probability by a third party. In contrast, by employing a bonus contract, the *Principal* offers the *Agent* an unconditional wage and the *Principal* can propose a non-binding bonus payment, which she can pay at her discretion upon observing the *Agent*'s effort choice. In both contracts, the *Agents*' effort choice is costly, but positively impacts the *Principals*' payoff. The main conclusion of the paper is that *Principals* reward *Agents*' efforts in a reciprocal manner and this promotes a more efficient outcome than the subgame perfect prediction.

The purpose of this paper is twofold. First, we conduct a conceptual replication of FKS in a *Replication* study, designed to re-examine the almost decade old conclusions of FKS.[2] Following FKS, *Principals* offer a contract, *Agents* exert an effort and then *Principals* can pay a bonus at their discretion. The experiment consists of ten periods with perfect stranger matching, removing any possibility of reputation building. In order to test the robustness of FKS' results, we make one small design modification by removing the *Principals'* ability to select incentive contracts. As the overwhelming number of contracts selected in FKS are incomplete bonus contracts (between 80% and 96% of choices in a given period of the experiment), if reciprocal motives are as strong as previously suggested, there is no reason to believe that giving *Principals* the

[1] The study is published in *Econometrica* and has over 400 citations.

[2] The definition of conceptual replication is taken from Crandall & Sherman (2016). The differences between *exact* and conceptual replications lie in how the experiments are implemented: a conceptual replication uses a different procedure to the targeted original study, whereas an exact replication follows the procedure of the targeted original study exactly.

opportunity to select between complete and incomplete contracts should unduly alter their behaviour. Our experiment is also computerised using zTree (Fischbacher, 2007), in contrast to FKS, which was conducted using pen and paper.

Second, we conduct an *Identity* study, where we experimentally induce subjects into artificial social groups. This is done to allow us to examine how a sense of group identity (Akerlof & Kranton, 2000) might affect reciprocal behaviour. Following the enhanced group paradigm of Chen & Chen (2011) and Chen & Li (2009), before subjects are assigned their role as *Principal* or *Agent*, they are randomly and anonymously assigned to one of two arbitrary groups, the *Red* or *Green* group. They then complete a task in which they can communicate with members of their group in order to identify six abstract paintings. Based on the results of Chen & Li (2009), a common sense of identity induced in this manner is hypothesised to increase reciprocity in 'in-group' interactions, or when they are matched with members of their own group, but decrease reciprocity in 'out-group' interactions. Previous studies have examined the effects a common sense of identity has on behaviour in public goods games (Drouvelis & Nosenzo, 2013), trust games (Falk & Zehnder, 2013), prisoners' dilemma games (Goette *et al.*, 2006) and dictator games (Whitt & Wilson, 2007), with all reporting significant in–group favouritism. We contribute to this literature by analysing the effect of a common identity in a gift–exchange game.

We report a number of findings. First, the results from the *Replication* study support the main conclusions of FKS. However, we find that reciprocity is not as strong as previously reported. In contrast to FKS, an overwhelming majority of subjects in the role of the *Principal* exhibit purely selfish behaviour. Where FKS report only 21% of *Principal* decisions as being selfish, we report 60%. Our experiment supports their results qualitatively, but not quantitatively. On average, we find reciprocity to be around 40% smaller than the original study, and significantly different. Although we cannot conclusively say why this is the case, we conclude that it is likely to be a consequence of procedural and subject pool differences, as discussed in Section 2.4.

Second, we find little evidence that a common sense of identity has any significant affect on behaviour: no significant differences are reported between choices in the identity and replication treatments. Further, both *Principals* and *Agents* behave *identically* towards both in–group and out–group members. Although this finding contrasts with much of the published literature, additional analysis reveals the induced identity may have interacted with the framing of the experiment, with *Agents* found to be more greatly attached to their group than *Principals*. This finding, that identities are multifaceted and may interact in ambiguous ways, is in line with the recent homophily literature (Charness *et al.*, 2014).

We make an additional contribution by estimating inequity aversion utility parameters structurally. We do this using both the original FKS data and the data from both our studies. This is done in order to compare parameter estimates between studies, but also to examine if these behavioural parameters are group–contingent (Chen & Li, 2009), or vary between in–group and out–group interactions. In line with the reduced form results, parameter estimates are significantly different between the *Replication* study and FKS, with the *Principals* in FKS found to be more inequity averse. No evidence that in–group or out–group identities affect *Principals'* preference parameters is reported.

The remainder of this paper is organized as follows. Section 2.2 reports the experimental design and procedure. Section 2.3 provides some brief behavioural predictions. Section 2.4 presents the results, and Section 2.5 presents structural estimates. Section 2.6 concludes.

2.2 Experimental design

The experiment was designed to examine the reproducibility of the reciprocity results of FKS and examine if a common, artificially induced identity could be used to foster reciprocal behaviour. Both studies followed the precise framing, parametrisation, session size and payoff structure of the original study by FKS in order to maximise comparability.

2.2.1 Replication study

Twenty subjects were recruited to each session, and were randomly and anonymously assigned to the role of either *Principal* or *Agent*. There were 10 subjects assigned to each role, which was maintained for the entire session. Each *Principal* was matched to each *Agent* exactly once, in a random order. Subjects completed 10 experimental periods per session and each period incorporated fully random, perfect-stranger matching, in order to remove the possibility for reputation building. Subjects received feedback about the outcome of each period and were paid for each period, as in FKS.

Each period consists of three stages. At $t = 1$, the *Principal* offers a contract to the *Agent* or not. If not, the game ends and both players receive zero payoffs. If she

Stage	*Principal*	*Agent*
$t = 1$	Offers a contract, $(w,\, e^*,\, b^*)$.	
$t = 2$		Observes the contract, $(w,\, e^*,\, b^*)$, chooses $e \in [0, .., 10]$.
$t = 3$	Observes effort, e, chooses $b \in [0, .., 100]$.	

Table 2.1: Timing of the Bonus Contract

does offer a contract, she specifies a wage, w, demands an effort from the *Agent*, e^*, and announces a bonus payment she could pay at $t = 3$, b^*. At $t = 2$, the *Agent* observes the contract and must choose an effort level, e. This effort can be different to demanded effort, e^*. Although effort is costly to the *Agent*, incurring him a cost $c(e)$, where $c(e) = \frac{e^2}{5}$, his effort choice increases the *Principal's* payoff by a factor of 10. At $t = 3$, the *Principal* observes e, and she decides on a bonus payment, b, which can differ to b^*.

Payoffs for the *Principal*, π_P, and payoffs for the *Agent*, π_A, are

$$\pi_P = 10e - w - b \tag{2.1}$$

$$\pi_A = w - c(e) + b \tag{2.2}$$

Following FKS, the choice sets were limited to integer values, with $e \in [0, 10]$, $w \in [0, 100]$, $e^* \in [1, 10]$, $b^* \in [0, 100]$ and $b \in [0, 100]$.[3] Table 2.1 outlines the timing of each period. Table 2.2 outlines the effort–costs faced by the *Agents* for each possible effort level.

The sub-game perfect equilibrium of this game is $b = 0$, $e = 1$, and $w = 0$; e^* and b^* can take any value, and *Agents* are assumed to treat this communication as cheap talk. Further Nash equilibria exist where $w \leq 10$, and $e = 1$. In Section 2.3 we briefly describe the behavioural predictions from the inequity aversion model of Fehr & Schmidt (1999), as in FKS.

Effort, e	1	2	3	4	5	6	7	8	9	10
Effort-Cost, $c(e)$	0	1	2	4	6	8	10	13	16	20

Table 2.2: The Agents' Effort-Cost Table

[3]When $e = 0$, the *Agent* rejects the contract, the game ends and both players receive a payoff of 0.

In total, we ran 4 sessions with 20 subjects in each, who all completed 10 periods. This gives a total of 400 contract offers from 40 *Principals*; this compares favourably to FKS, who ran 2 sessions, and collected 198 observations from 20 *Principals*.

2.2.2 Identity study

The identity study followed the replication study exactly, but added an additional stage prior to the 10 periods of the bonus contract being completed. Following the enhanced group procedure of Chen & Chen (2011), participants were first randomly and anonymously assigned to one of two groups, the *Red* or the *Green* group. Group sizes were equal to control for potential minority-group effects (Tsutsui & Zizzo, 2014). Subjects then completed a group problem solving task.

The problem solving task required subjects to first privately study six paintings, three by Paul Klee and three by Wassily Kandinsky. Subjects were then shown four additional paintings which they had to try to identify within a five minute time limit, earning £0.80 for each correct answer. To enhance the sense of group identity, subjects could anonymously communicate with their own group using a chat program. No feedback was given on the number of correctly identified paintings until the end of the session, to control for any income effects. Subjects were never told which paintings they correctly identified, and were not aware of the nature of the task that would follow this procedure.[4] This procedure, as shown by Chen & Li (2009), is designed to make group identity salient to the subjects and is more likely to influence behaviour than the minimal or near–minimal group paradigms.

Once completed, subjects then played 10 periods of the bonus contract, as outlined in Section 2.2.1. Half the subjects assigned to the role of *Principal* were *Red*, the other half *Green*; the same for *Agents*. As each *Principal* is matched to each *Agent* exactly once in a random order, we always observe half in–group (*Red/Red* and *Green/Green*) and half out–group (*Red/Green Green/Red*) interactions per subject, per session. We conducted 4 sessions in total.[5]

At the end of each session, following the literature (Chen & Li, 2009; Ioannou *et al.*, 2015; Turner, 1975), we asked participants to answer the following question: *"On a scale from 1 to 10, please rate how closely attached you felt to your own group throughout the*

[4]These pictures, and all experimental materials, are given in Appendix A.1.

[5]Due to subjects failing to show up, we completed 2 sessions with 20 subjects and 2 sessions with 16 subjects. The sessions with 16 subjects only completed 8 periods of the bonus game to avoid re-matching.

experiment", with 1 meaning *"not at all"* and 10 meaning *"very much so"*. We use this measure of group *attachment* to provide some indication of how successfully identity was induced. This measure will also allow us to examine how identity might interact with the subjects' role, as either *Principal* or *Agent*.

2.2.3 Summary

Study	*Group Task*	*Bonus Contract*	*Sessions*	*Matching*	*Subjects*
Replication		✓	4	No Identity	80
Identity	✓	✓	4	In-group/Out-group	72

Table 2.3: Summary of Studies

Table 2.3 summarizes the two studies and the differences between them. Before the sessions began, subjects were read instructions aloud and were required to complete exercises to ensure that they understood the task (see Appendix A.1 for all experimental materials). All sessions were conducted at BEEL (Birmingham Experimental Economics Laboratory) at the University of Birmingham UK during the Autumn Term of 2013. The sessions lasted for around 90 minutes, and subjects earned £13.90 on average; maximum earnings were £29, whilst the lowest were £5.20. The sessions were conducted using zTree (Fischbacher, 2007), which contrasts to Fehr *et al.* (2007), who conducted their experiments using pen and paper.

2.3 Behavioural predictions

In this section we examine the predictions of the outcome based model of inequity aversion (Fehr & Schmidt, 1999). Due to the sequential nature of the game, behaviour in Stages $t = 1$ and $t = 2$ is unlikely to representative of any underlying preference, so focus is placed on the *Principals*' choice of bonus payment in the final stage, Stage $t = 3$.

If the *Principal*, P, is assumed to be inequity averse, her utility is assumed to be a function $u : \pi \rightarrow \mathbb{R}$, where $\pi = (\pi_P, \pi_A)$ denotes the vector of monetary payoffs. As defined in Section 2.2, the *Principal's* payoff is denoted as π_P and the *Agent*'s as π_A. The utility function for the *Principal*, u_P, is assumed to take the following form,

$$u_P(\pi) = \pi_P - \alpha_P \max\{\pi_A - \pi_P, 0\} - \beta_P \max\{\pi_P - \pi_A, 0\}. \tag{2.3}$$

If the *Principal* is purely selfish, then $\beta_P = \alpha_P = 0$. However, if $\beta_P > 0$ and $\alpha_P > 0$, the *Principal* will seek to equalise *Principal* and *Agent* payoffs, $\pi_P = \pi_A$. This gives the following utility maximising choice of bonus payment,

$$b = \max\left\{5e - w + \frac{e^2}{10}, 0\right\} \tag{2.4}$$

when $\pi_P > \pi_A$, zero otherwise. As shown, bonus payments are predicted to be increasing in effort and decreasing in wages. In Section 2.5 we structurally estimate the behavioural parameters of the proposed utility function using the experimental data.

The recent work of Chen & Li (2009) suggests that the *Principals*' preference parameters, α_P and β_P, are functions of the group identity of the individual with whom they are interacting, or are group–contingent. Let $g \in \{I, O\}$ be an indicator function, with $g = I$ when the *Agent* is In–group, and $g = O$ when the *Agent* is out–group. We can then write each utility parameter as functions of identity: β_P^g and α_P^g. Given the findings of Chen & Li (2009), when estimating the parameters we expect to find $\beta_P^I > \beta_P^O$ and $\alpha_P^I > \alpha_P^O$, suggesting Principals' have a greater preference for equal payoffs when faced with an in–group *Agent*.

2.4 Results

In this section, we first compare the *Replication* study to the results of FKS. Data reported from FKS is taken from the bonus contracts of the *Bonus–Incentive* treatment, which was downloaded through the *Econometrica* website.[6] We then examine the data from the *Identity* study and analyse it for any effects stemming from an induced group identity.

A number of common features are present throughout: where non-parametric tests are given, the *p*-value and test used are both presented in parentheses. Unless stated otherwise, these tests are always two sided and the null hypothesis is that there is no effect. Following FKS, all comparisons are made using observations at the individual level.[7]

[6] We were able to estimate all reported results in the paper exactly.

[7] The results do not change significantly if session level observations are used instead.

2.4.1 Replication study

Table 2.4 reports summary statistics from the *Replication* study, alongside those from FKS to allow for ease of comparison. Table 2.4 presents Period 1 averages, and pooled Period averages, from both studies. Figure 2.1 plots the evolution of the variables over time, for both the *Replication* study (Figure 2.1a) and for FKS (Figure 2.1b).

Study		*Replication*		*FKS*	
		Period 1	*Average*	*Period 1*	*Average*
Wage	w	13.35 (10.85)	14.6 (10.84)	16.7 (7.52)	15.2 (6.8)
Demanded Effort	e^*	6.59 (2.51)	7.1 (2.38)	6.3 (1.59)	6.7 (2.04)
Announced Bonus	b^*	12.49 (11.05)	21.9 (17.19)	21.4 (6.73)	25.1 (10.4)
Effort	e	5.08 (2.49)	3.6 (2.83)	5.55 (2.46)	5.3 (2.76)
Bonus	b	5.84 (6.12)	3.3 (6.31)	11.45 (8.37)	10.4 (10.9)
Contracts		37	376	20	198
Subjects		37	40	20	22

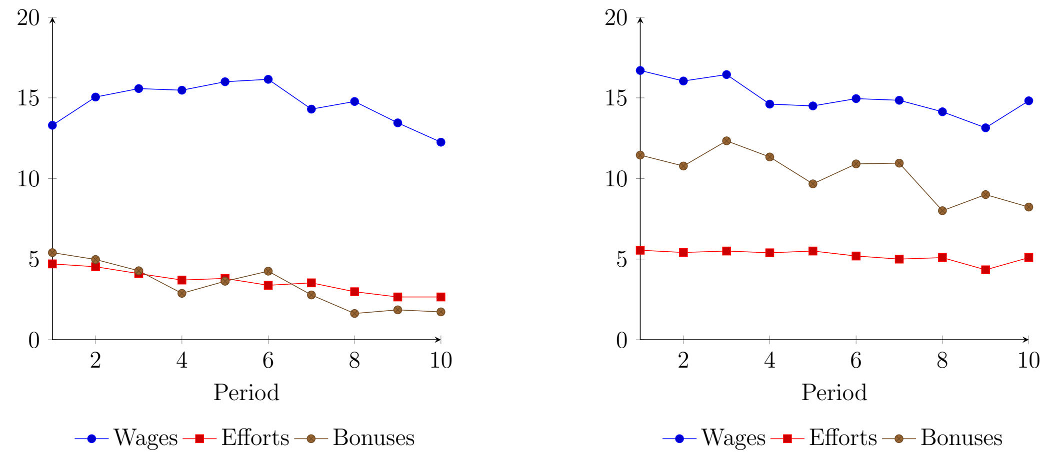
Note: Standard deviations in parentheses. FKS data taken from the bonus contracts of the *Bonus-Incentive* treatment.

Table 2.4: Summary Statistics

Result 1: *The behaviour of Agents in Period 1 is the same across studies. However, Principals' behaviour in Period 1 of the Replication study is significantly different to that observed by FKS.*

Support. In Period 1, we observe significant differences in wages, announced bonuses and bonus payments ($p < 0.01$, in all cases, Sign Tests). Demanded efforts in Period 1 and initial effort choices are equal across studies, and statistically indistinguishable ($p > 0.1$, in both cases, Sign Tests). The significant differences in bonus payments is easily discerned in Figure 2.1.

Result 2: *Principals reciprocate higher effort levels with larger bonus payments. However, reciprocal behaviour is significantly smaller in the Replication Study in comparison to FKS.*



(a) Replication Study

(b) FKS

Figure 2.1: Wages, Efforts and Bonuses by Period

Support. On average, wages, announced bonuses and demanded efforts are not significantly different at the 5% level ($p = 0.08$, $p = 0.15$, $p = 0.43$, Sign Tests), although wages are significantly different at the 10% level. On average, contract offers are reasonably similar between studies. However, average efforts and bonus payments are significantly different between studies at the 1% level ($p < 0.001$, in both cases, Sign Tests). Figure 2.2 plots the distributions of bonus payments; FKS report less than 40% of all bonus payments being zero, whereas we report 60%. This difference is statistically significant ($p < 0.05$, Binomial Test).

To support the non–parametrics, Table 2.5 presents a number of OLS and Tobit regressions. Models *i*, *ii* and *iii* present estimates using data from the *Replication* study that follow the analysis of FKS exactly: in each model the dependant variable is the bonus payment and, guided by the behavioural predictions in Section 2.3, we include wages, demanded efforts, announced bonuses and efforts as explanatory variables. Models *iv*, *v* and *vi* present estimates taken from FKS.

From models *i*, *ii* and *iii*, the estimated coefficient on effort is positive and highly significant and is robust across models. However, the estimated coefficient on effort is around half the size of that estimated by FKS, and in each model we can reject the null hypothesis that the coefficient on effort is the same as the corresponding FKS estimate ($p < 0.001$ in all cases, χ^2 Test). Reciprocity is not as large as reported in FKS.

Additional differences are also reported in Table 2.5, such as the negative and significant correlation between announced bonuses on bonus payments ($p < 0.01$, model

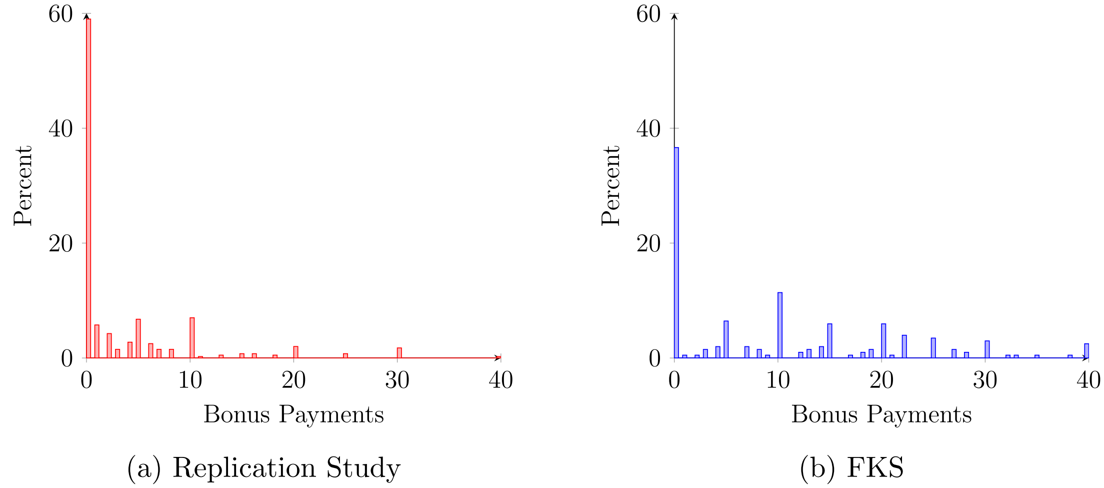


Figure 2.2: Distributions of Bonus Payments

i and *iii*, χ^2 Tests). This contrasts with FKS, who find that announced and actual bonuses correlate positively and insignificantly, rather than negatively.

Insight into the potential source of the differences between studies can be gained from Results 1 and 2. First, the contract offers of *Principals* are incredibly similar between studies, and the initial behaviour of *Agents* is identical to that observed in FKS. This suggests that, at least initially, subjects in the role of an *Agent*, in both studies, have similar beliefs about the behaviour of *Principals* in Period 1. However, the largest difference between studies is observed in the bonus payment, the decision that reveals the *Principals'* preference for reciprocity and the *least* strategic choice within this setting. The study differential in bonus payments begins in Period 1 and persists until the final Period, which suggests that *Principals* are simply less reciprocal in the *Replication* study than in FKS.

One explanation for the difference between studies might be found in the differences in the way the experiments were conducted. Here we computerised experiments using zTree, whereas FKS utilised pen and paper, the former perhaps providing greater abstraction away from human interaction. In line with the recent findings of Malmendier *et al.* (2014), it may be that this resulted in the 'external' motivations of subjects being different between studies. For example, as the subjects in FKS interacted with the experimenter more often due to the pen and paper procedure, they may have perceived greater social pressures that encouraged them to behave more reciprocally.

Alternative explanations might be due to the fact that this study was conducted in

Study:	*Replication*			*FKS*		
	(i)	*(ii)*	*(iii)*	*(iv)*	*(v)*	*(vi)*
Effort	1.549***	1.549***	3.296***	2.86***	2.86***	5.54***
	(0.153)	(0.206)	(0.369)	(0.20)	(0.33)	(0.55)
Demanded Effort	0.135	0.135	-0.182	0.33	0.33	-0.59
	(0.106)	(0.152)	(0.317)	(0.38)	(0.46)	(0.77)
Wage	-0.056**	-0.056	-0.136	-0.3***	-0.3*	-0.54**
	(0.028)	(0.044)	(0.093)	(0.1)	(0.17)	(0.24)
Announced Bonus	-0.039***	-0.039*	-0.181***	0.12*	0.12	0.11
	(0.013)	(0.021)	(0.043)	(0.06)	(0.08)	(0.11)
Constant	-1.623**	-1.623	-7.553***	-5.58***	-5.58***	-14.55***
	(0.744)	(1.062)	(2.549)	(1.88)	(2.59)	(3.68)
Observations	376	376	376	198	198	198

Note: The dependant variable is the bonus payment. ***, ** and * denote significance at the 1%, 5% and 10% levels. Standard errors in parentheses. Models (*i*) and (*iv*) are OLS regressions with robust standard errors. Model (*ii*) and (*v*) are OLS regressions with robust standard errors clustered at the individual level. Model (*iii*) and (*vi*) are Tobit regressions with robust standard errors clustered at the individual level, lower censored at 0. FKS estimates are taken from authors own estimates, but are also available in regressions (1), (2) and (3) from (Fehr *et al.*, 2007, pp 139).

Table 2.5: Determinants of Bonus Payments

the UK, and FKS was conducted in Germany. They were also conducted around eight years apart. Therefore, we cannot rule out cultural factors stemming from nationality or having developed over time, as driving the differences between studies. Finally, the *Replication* study also differs in that *Principals* are unable to offer complete 'incentive compatible' contracts to the *Agents*, as in FKS. Although an unlikely explanation, we cannot rule out the possibility that the removal of this design feature adversely affected *Principals*' reciprocity.

2.4.2 Identity study

We now present the results from the *Identity* study. Table 2.6 presents summary statistics, disaggregated by *In–group* and *Out–group* interactions. Similar to the summary statistics in Table 2.4, we present both Period 1 averages and averages across periods.

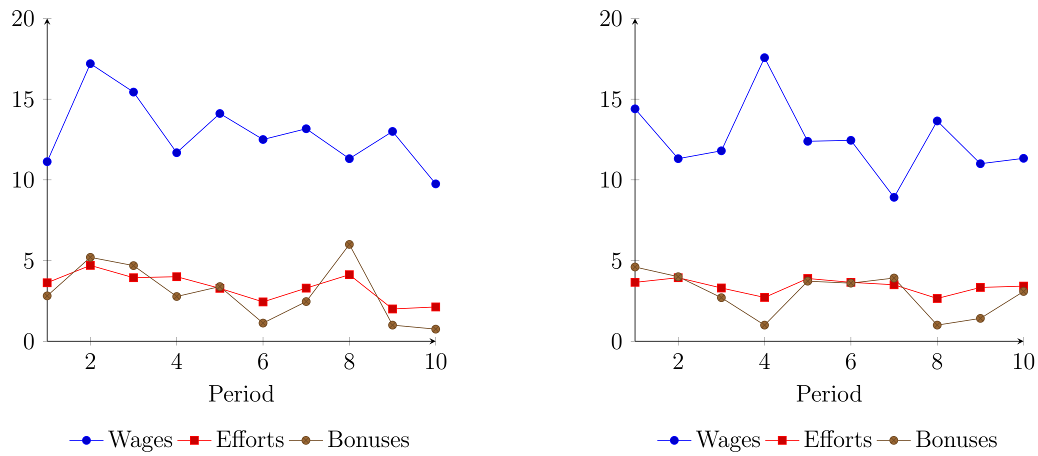
Result 3: *A common induced sense of identity has no effect on the Principals' or Agents' behaviour.*

Support. Comparing first period contract offers, no significant differences are re-

Matching:		*In–group*		*Out–group*	
		Period 1	*Average*	*Period 1*	*Average*
Wage	w	11.2	13.37	14.4	12.94
		(7.04)	(8.51)	(12.25)	(8.93)
Demanded Effort	e^*	6.29	6.48	6.6	6.48
		(2.3)	(2.17)	(2.21)	(2.13)
Announced Bonus	b^*	12.14	18.44	12.7	18.68
		(12.62)	(13.83)	(12.6)	(13.8)
Effort	e	4.14	3.73	3.65	3.56
		(1.75)	(1.67)	(2.25)	(1.61)
Bonus	b	3.21	3.53	4.6	2.98
		(3.58)	(4.41)	(6.21)	(4.04)

Note: Standard deviations in parentheses.

Table 2.6: Summary Statistics



(a) In–group (b) Out–group

Figure 2.3: Wages, Efforts and Bonuses by Period

ported when comparing in–group wages, demanded efforts or announced bonuses to those offered to the out–group ($p > 0.1$ in all cases, Robust Rank Order Tests). This is despite out–group *Agents* being offered higher wages in Period 1 in comparison to in–group *Agents*. Similar results are reported for efforts and bonus payments, with no significant differences reported ($p > 0.1$ in both cases, Robust Rank Order Tests). On average, no significant differences are reported in contract offers, effort choices or bonus payments ($p > 0.1$ in all cases, Robust Rank Order Tests). The similarities are highlighted graphically in Figure 2.3.

Further support for Result 3 is provided in Table 2.7, which presents the estimates from a number of regressions. In each model, the dependent variable is the bonus payment, and we include identical explanatory variables as those outlined in Table 2.5, with the addition of a dummy controlling for *Out–group* interactions (1 if yes, 0 otherwise) and the interaction between effort and the *Out–group* dummy. In–group interactions are taken as the baseline.

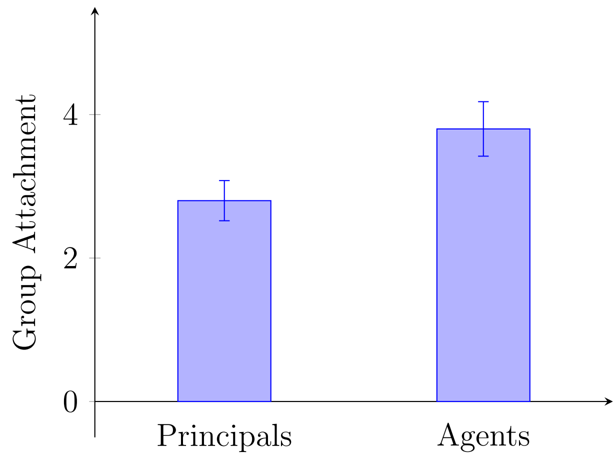
Table 2.7 outlines how the coefficient on the out–group dummy is never significant at conventional levels ($p > 0.1$, in all cases, χ^2 Tests). The interaction between effort and the *Out–group* dummy is also found to be insignificant ($p > 0.1$, in all cases, χ^2 Tests).

Study:	*Identity*		
	(*i*)	(*ii*)	(*iii*)
Effort	1.304***	1.303***	2.83***
	(0.245)	(0.316)	(0.604)
Effort × *Out–group*	-0.175	-0.175	-0.07
	(0.357)	(0.397)	(0.577)
Demanded Effort	-0.181	-0.181	-0.642
	(0.147)	(0.182)	(0.431)
Wage	-0.019	-0.019	-0.137
	(0.029)	(0.047)	(0.099)
Announced Bonus	0.05	0.05	-0.052
	(0.037)	0.062	(0.119)
Out–group	0.466	0.466	0.111
	(0.869)	(0.999)	(0.194)
Constant	-1.623**	-0.919	-5.613**
	(0.744)	(0.945)	(2.666)
Observations	309	309	309

Note: ***, ** and * denote significance at the 1%, 5% and 10% levels. The dependent variable is the bonus payment. Standard errors in parentheses. Model (*i*) is an OLS regression with robust standard errors. Model (*ii*) is an OLS regression with robust standard errors clustered at the individual level. Model (*iii*) is a Tobit regression with robust standard errors clustered at the individual level.

Table 2.7: Determinants of Bonus Payments - Identity Study

Result 3 suggests that an induced group identity has no effect on behaviour. This is in contrast to many papers that report significant differences resulting from random assignment to arbitrary groups (Chen & Li, 2009; Chen & Chen, 2011; Drouvelis & Nosenzo, 2013; Eckel & Grossman, 2005; Ioannou *et al.*, 2015). One explanation for this is that the experimental identity manipulation was not strong enough, or that group identities may have interacted with the framing of the experiment. To examine this, we consider the subjects' self–reported group *attachment*, as is standard in the literature. These measures are presented in Figure 2.4, disaggregated by the experimental role of the subjects.



Note: Vertical bars represent 95% confidence intervals.

Figure 2.4: Average Self–Reported Group Attachment

It is clear from Figure 2.4 that *Agents* report feeling closer to their group than do *Principals*. The difference is also found to be statistically significant ($p < 0.01$, Robust Rank Order Test). This may explain why *Principals* do not treat in–group *Agents* better than out–group *Agents*: they do not feel as closely attached to their group. Rather than the identity manipulation being unsuccessful, it appears as if the non–neutral framing of the experiment interacted with the group identity of the subjects. This finding would be consistent with the recent results of Charness *et al.* (2014), who find that an induced group identity is not enough to overcome an unanticipated 'status' identity that was potentially induced by differences in subjects' productivity. In our study, although *Principals* may have identified with their group they may *not* have identified with subjects in the role of *Agent* perhaps as a consequence of a perceived status. This status effect likely stems from the fact that *Principals* earned significantly more on average than *Agents* (£17 versus £7 respectively).

2.5 Structural estimates

In this section, we conduct a structural analysis in order to link our results to behavioural theory. Specifically, we estimate the behavioural parameters proposed in the model of inequity aversion in Section 2.3. However, we are only able to estimate the advantageous inequality parameter, β, from Equation 2.3, as in every observed contract, before the bonus payment is made, the *Principals*' payoff is *always* greater than the *Agents*'. Implicit in the analysis is the assumption that *Principals* are making their decision in isolation, or that subjects are 'narrow bracketing'. Thus we assume they are not taking into account the total income of the *Agent* or themselves. It has been suggested that this may result in greater estimates of 'fair' behaviour than if they were assumed to 'bracket broadly', and considered total incomes (Read *et al.*, 1999).

For a given bonus payment, $b = x$, when $\pi_P > \pi_A$, each *Principal* derives utility

$$u_P(b = x) = 10e - w - x - \beta\left(10e - w - x - (w - c(e) + x)\right). \tag{2.5}$$

Following the estimation strategy of Chen & Li (2009), to estimate β, we assume *Principals* select a bonus payment, $b = x$, with the following probability

$$Pr(b = x; \beta, \lambda) = \frac{e^{\lambda u_P(b=x)}}{\sum\limits_{i=0}^{100} e^{\lambda u_P(b=i)}} \tag{2.6}$$

where u_P is the utility function specified in Equation 2.5 and λ is a noise parameter. We assume a random utility model, rather than a random preference model, due to the assumed linearity of the utility function (Fehr & Schmidt, 1999). When $\lambda \to 0$, the *Principal* makes perfectly random choices over the set of possible bonus payments, $b \in \{0, 1, ..., 100\}$. As $\lambda \to \infty$ the *Principal* makes a perfectly rational choice, given her utility function. We estimate β and λ using maximum–likelihood estimation, with the results presented in Table 2.8 under the *Replication Study* heading. For comparison purposes we provide the inequality aversion parameter estimates using the FKS data, which have not previously been estimated. These estimates are presented in Table 2.8 under the *FKS* heading.

In addition, following the behavioural literature, we specify the *Principals*' advantageous inequality parameter, β, as a function of in–group and out–group identity to allow us to test for group–contingent social preferences (Chen & Li, 2009; Chen & Chen, 2011). We specify β as a function,

$$\beta = \beta(1 + aI + dO) \tag{2.7}$$

where I is a dummy variable that takes a value of 1 if the *Agent* is In–group, and O takes a value of 1 if the *Agent* is Out–group, both 0 otherwise. Therefore, parameters a and d capture the additional utility weight placed on the *Agents*' payoff when they are either in–group, or out–group. We interpret β, which captures the preference parameter estimated from the *Replication* study, as being the social preference parameter where identities are group–neutral (Chen & Li, 2009). These estimates are presented in Table 2.8 under the *Identity Study* heading.

Model: *Parameter*	*Replication Study*	*FKS*	*Identity Study*
β	0.244*** (0.016)	0.284*** (0.013)	0.251*** (0.121)
λ	0.512*** (0.031)	0.216*** (0.021)	0.527*** (0.026)
a			-0.01 (0.106)
d			-0.05 (0.121)
Observations	400	202	728
Log–Likelihood	-936.8	-608.7	-1683.6

Note: ***, ** and * denote significance at the 1%, 5% and 10% levels. Standard errors clustered at the individual level are given in parentheses. Estimates are of the parameters of the Fehr & Schmidt (1999) model of inequality aversion.

Table 2.8: Structural Parameter Estimates

First, the estimates of β and λ in all models are positive, and highly significant ($p < 0.001$ in all cases, χ^2 Test). However, comparing the results between the *Replication* study and FKS, reveals that β is significantly different across studies ($p = 0.003$, χ^2 Test), as is λ ($p < 0.001$, χ^2 Test). Although β is estimated to be around 16% larger in FKS, suggesting greater average levels of inequality aversion, the differences in λ suggests FKS subjects randomise over the set of bonus payments to a greater extent than those in the *Replication* study. Parameter λ estimated from the FKS data is less than 50% the size of λ estimated using the *Replication* study data. The estimates of β also support the reduced form conclusions in Section 2.4.[8]

[8]Our estimates of β, from both the *Replication* study and FKS, are in line with the median values

The *Identity* study estimates are also in line with the results in Section 2.4: in–group and out–group identities have no significant affect on the *Principals'* preferences, with both a and d not found to be significantly different from zero ($p > 0.1$ in both cases, χ^2 Tests).

2.6 Conclusion

We report the results of experiments designed to replicate the results of Fehr *et al.* (2007). Although we find evidence that supports the qualitative evidence of Fehr *et al.* (2007), we find that subjects are significantly less reciprocal than subjects in the original study, with reciprocity estimated to be around 40% smaller. Structural estimates reveal subjects preferences to be significantly smaller in our study in comparison to the original.

We also find that an induced group identity has little influence on reciprocity in this particular laboratory setting. The data suggests that this is a result of an interaction between the identity manipulation and the framing of the experiment; *Agents* feel more closely attached to their group than the *Principals*, potentially due to a status effect. Although this appears to go against some of the literature on identity, it is in line with the homophily literature. Our result outlines how identity is likely to be complex and multifaceted even when induced (Charness *et al.*, 2014; Chen *et al.*, 2014), and shows how it is not always clear which sense of identity will be made salient. However, we acknowledge that we cannot rule out the possibility that the identity manipulation was unsuccessful.

In conclusion, we find that subjects are less reciprocal than previously reported. This is most likely attributable to subject pool differences, and slight differences in the experimental procedure. Specifically, we believe a move from a pen and paper based procedure to one conducted on a computer may have induced greater 'external motivations'. Pen and paper based experiments involve far more interaction with the experimenter, which may induce subjects to behave more prosocially than they would otherwise. This conclusion is in line with the conclusions of Malmendier *et al.* (2014), that external motivations likely play an important role in shaping reciprocal behaviour. As recently highlighted in the replication studies of Camerer *et al.* (2016), only further replication of experimental findings can shed light on their robustness.

---

of β that are typically estimated and assumed. See, for example, Blanco *et al.* (2011).

Chapter 3

Discrimination in a Deprived Neighbourhood: A Field Experiment[*]

"They're not coming to this country if I'm president"
— Donald J. Trump, 2015

3.1 Introduction

Donald Trump, the Republican Party presidential candidate, was recently branded a bigot and a racist for his proposed policy of banning Muslims from entering the United States (Milibank, 2015). His comments were described as "hate speech," and the United Kingdom's House of Commons debated banning him from entering the country after a petition calling for his ban was signed by 500,000 people.[1] Similar anti–Muslim sentiment was at the heart of the recent election of the London Mayor, Sadiq Kahn and the 'Vote Leave' campaign during the British referendum on EU membership. For example, those standing in opposition to Sadiq Kahn made a number of allegations linking him to Muslim extremist groups, and Vote Leave directed significant attention

[1] The petition, parliamentary debate and discussion can be viewed through the UK's Parliamentary Petition website: https://petition.parliament.uk/petitions/114003.

[*] The field experiment reported in this chapter was conducted in collaboration with Brit Grosskopf.

to potential increased immigration from Muslim countries in order to promote anti–EU sentiment.[2]

The similarity between the campaigns is likely not a coincidence, as supporters of both Trump and Vote Leave appear to share similar views and are from a similar demographic. As highlighted by a number of pre and post–referendum polls, voters most likely to have supported Vote Leave are white, have low levels of education and earn low incomes.[3] The same appears to be true for the supporters of Donald Trump (Thompson, 2016). Further, surveys suggest that 30% of British voters would support policies that reduced the Muslim population (Townsend, 2012), and around 27% believe Islam is incompatible with the values of British democracy. This latter view is found to be correlated with respondents incomes and level of education (Page, 2009).

However, little is known about *why* individuals from this particular demographic might support policies that discriminate against those with Islamic heritage. This is largely because there is little research into the behaviour of the poorer, less educated population, with previous work into discrimination against Muslims typically focusing on better educated and more wealthy individuals that are less likely to be supporters of such policies. For example, Ahmed (2010*b*) reports evidence of discrimination in trust and dictator games against non–Europeans in an experiment utilising Swedish university students. Ahmed & Hammarstedt (2008) find that landlords in the Swedish housing market discriminate against tenants with Muslim sounding names in comparison to 'native' names. Ahmed *et al.* (2009) find that small business sellers are less likely to contact potential buyers if they have Muslim sounding names. Kaas & Manger (2012) report similar evidence of discrimination by hiring committess in the German labour market towards those with Turkish sounding names. Booth *et al.* (2012) report comparable results from the Australian labour market against those with names of Middle Eastern origin. However, the strategic nature of the interactions in these studies means that

[2] A number of newspaper articles demonstrate this point. For example, Hinsliff (2016) reports on Sadiq Kahn's competitors trying to link him to an ISIS sympathiser. The former UKIP party leader, and prominent figure in the Vote Leave campaign, Nigel Farage, posed with a billboard poster of Muslim refugees with the slogan, 'Breaking Point' (Wright, 2016), and Vote Leave 'appealed to prejudice' by claiming Turkey was about to join the EU and that its citizens posed a threat to UK national security (Boffey & Helm, 2016).

[3] Three separate studies support this. The *Lord Ashcroft Poll* information on voter demographics: http://lordashcroftpolls.com/2016/06/how-the-united-kingdom-voted-and-why/; the *Telegraph* newspaper poll: http://www.telegraph.co.uk/news/2016 /06/22/eu-referendum-which-type-of-person-wants-to-leave-and-who-will-b/; and a *Guardian* report: https://www.theguardian.com/news/datablog/2016/jun/24/the-areas-and-demographics-where-the-brexit-vote-was-won.

individuals may be using ethnic stereotypes in order to inform their decisions, making preference (Becker, 1971) and statistical (Phelps, 1972) explanations of discrimination difficult to disentangle.

In contrast to field studies, lab experiments provide a more tightly controlled environment in which statistical and taste–based explanations of discrimination can be more easily parsed. Experimenters have employed a range of strategic and non–strategic games in order to distinguish between the two competing explanations, but also to develop new behavioural theories of discrimination. Recent studies suggest that individuals' social preferences are group–contingent, or that the extent to which individuals care about others depends on the degree to which they identify with them (Akerlof & Kranton, 2000; Chen & Li, 2009). Individuals have been found to behave more charitably (Chen & Li, 2009; Whitt & Wilson, 2007), cooperatively (Drouvelis & Nosenzo, 2013; Falk & Zehnder, 2013; Goette *et al.*, 2006; Ruffle & Sosis, 2006) and coordinate more efficiently (Chen & Chen, 2011) when interacting with those they perceive as the 'in–group', in comparison to the out–group. A stylised interpretation of these findings is that individuals exhibit favouritism towards the in–group, rather than negativity towards the out–group, what Bernhard *et al.* (2006) call *parochialism*. As noted by Becker (1971), this is an important behavioural distinction, and as suggested by Ahmed (2007), many studies that conclude that individuals exhibit in–group favouritism omit a treatment where interactions are 'group neutral' (Falk & Zehnder, 2013; Fershtman & Gneezy, 2001; Goette *et al.*, 2006; Ruffle & Sosis, 2006; Whitt & Wilson, 2007), and so are unable to parse behaviour into in–group favouritism or out–group negativity.

The purpose of this paper is to shed light on the potential motives underpinning the behaviour of a previously unstudied demographic, poor and less educated white-British individuals, and determine if their support for anti–Muslim policies could be a result of taste–based discrimination (Becker, 1971). This is achieved using a door–to–door field experiment conducted in the poorest areas in England, two housing estates in the town of Rochdale. The residents of these estates were chosen because they are highly representative of the demographic of interest: they are white, have received little formal education and have very low incomes. Many of those who took part in our experiment live on incomes as low as £57 per week, with some subjects living in neighbourhoods that endure unemployment rates of 100%.

Subjects were first asked to make distributional choices in Other–Other games, where they had to divide £10 between two anonymous individuals. Following the social psychology experiments of Turner (1978), this was done in order to make their ethnicity salient. They then played a single Dictator Game, dividing £10 between themselves and a receiver. Similar to Fershtman & Gneezy (2001), and the lost letter experiments of Ahmed (2010*a*), the 3×1 between–subject design subtly varies the ethnicity of the individuals that subjects are allocating money to by providing them with surnames taken from the local electoral register.[4] The surnames were categorised into either *English* or *Muslim* ethnic origin using the taxonomy of Mateos *et al.* (2007). These allocation decisions are then compared to the decisions from an *Anonymous* baseline treatment, where the receiver's surname is withheld, allowing us to distinguish between in–group favouritism and out–group negativity. As we study a non–standard subject pool, the experiment satisfies the artefactual field experiment criteria of Harrison & List (2004).

In the Dictator Game, we find that subjects give around £5 to receivers with surnames of *English* origin, £2 to those with surnames of *Muslim* origin and £5 to those in the *Anonymous* treatment. Individuals with surnames of *Muslim* origin are treated worse than someone who is *Anonymous*, whilst those with surnames of *English* origin are not treated more favourably. Thus, in contrast to the conclusions of the majority of the literature we find no evidence of in–group favouritism, but instead report evidence of out–group negativity. This is a particularly interesting finding, as anti–immigration and nationalist groups often associated with this demographic employ slogans or names that focus on the in–group, and imply that they are promoting in–group favouritism, rather than supporting out–group negativity.[5]

To link our empirical analysis to behavioural theory, we assume that subjects give in the Dictator Game because they have social preferences, and that these preferences are group–contingent (Chen & Li, 2009). Structural parameter estimates reveal that social preferences are 87% smaller when subjects are giving to a receiver with a surname of *Muslim* origin in comparison to receivers in the *Anonymous* treatment and those with surnames of *English* origin. In Section 3.6 we conduct a counterfactual simulation using the preference parameters we estimate, and discuss how they could provide an explanation for our subjects' potential support of discriminatory policies. Our conclusions may also provide insight into the results of Ahmed & Hammarstedt (2008), Booth *et al.*

[4]Experiments into gender differences conducted by Holm (2000) preceed those of Fershtman & Gneezy (2001) and use a similar method, but utilise forenames rather than surnames in order to study gender.

[5]For example, consider the right–wing British organisation, *Britain First*, or nationalist policies promoted in the UK such as 'Buy British' or phrases such as 'British Jobs for British Workers'.

(2012), and the discrimination faced by Muslims reported in the press.[6]

In the Other–Other games we find that when allocating money between two individuals with surnames of contrasting ethnic origin, one *English* and one *Muslim*, subjects allocate around £1 more to the individual with the surname of *English* origin regardless of the ordering of the surnames. Although this result resonates with the minimal group literature (Turner, 1978; Chen & Li, 2009), we cannot determine if this behaviour is driven by in–group favouritism or out–group negativity.

This study makes a number of contributions. First, we contribute to the discrimination literature by examining the behaviour of a previously unstudied population. Second, we provide some of the first evidence into the potential motives underpinning the behaviour of this novel demographic. Finally, we provide a potential behavioural rationale for *why* these individuals might discriminate by modelling ethnic discrimination as being a consequence of group–contingent social preferences.

The remainder of this paper is organised as follows. Section 3.2 gives details of the study population, Section 3.3 outlines the experimental design, Section 3.4 outlines the results, Section 3.5 presents a structural model, Section 3.6 provides a counterfactual analysis, and Section 3.7 concludes.

3.2 Subject pool

The subjects in our experiment are drawn from two population areas, *Falinge* and *Kirkholt*, two housing estates situated in Rochdale, England, a town located in a wider region that has a recent history of ethnic tensions (Carter & Midlane, 2012; Syal & Topping, 2014). The extent of unemployment in these areas is most evident from the English Indices of Deprivation, a five yearly publication from the UK Government's Office of National Statistics (ONS). The report ranks small highly localised populations in terms of relative deprivation. At the time the study took place (2014) the *Falinge* and *Kirkholt* housing estates were ranked in the top 0.3% of the most income and employment deprived populations in the country. The *Falinge* estate was ranked first out of 32,482. It was determined to be the most deprived area in both these domains.[7]

[6] For example, Syal & Topping (2014) report on taxi customers in Rochdale requesting 'local' (white British) drivers over the phone, rather than Asian–Muslim drivers.

[7] In the latest 2015 report, other locations have become relatively more deprived. However, these locations are still ranked in the top 1% of the most income and employment deprived populations in the country.

Estate	*Census Area*	*No. Households*	*Claiming Benefit*‡	*No Education*†	*White*	*Deprivation Domain*◇ *Income*	*Employment*
Falinge	F1	177	85%	31%	40%	1st	1st
	F2	215	100%*	44%	60%		
Kirkholt	K1	132	75%	44%	82%	98th	50th
	K2	120	65%	31%	66%		
	K3	150	85%	36%	79%		
	K4	136	55%	36%	80%		

Source: Office of National Statistics, English Indices of Deprivation 2010, UK Census 2011 and own calculations. The Output Area codes used within the Census are removed for anonymity reasons. *The Census reports this figure as 120%, which is potentially attributed to fraudulent benefit claims. ‡Percent of population claiming out of work benefits. †Percent of population with no formal qualifications. ◇Income and employment deprivation ranks out of 32,482. Higher ranks imply greater levels of deprivation.

Table 3.1: Population Demographics

Table 3.1 provides information on the housing estates we study. Each estate is divided into different areas by the UK Census, and the Table outlines the number of households, the percentage of people without qualifications, the percentage of people out of work and simplified ethnic demographics for each of these Census Areas. It is evident that a significant number of residents are out of work, and a large proportion have not obtained any formal qualifications. Deprivation rankings in the income and employment domains are given in the final two columns.

Although a majority of residents from the two housing estates are white British nationals, Table 3.1 highlights that a large number of residents are from non-white minority ethnic groups. In one area in *Falinge* (area F1), non–white residents constitute a majority of the population (60%). In contrast, far fewer minority ethnic groups are present in the population of the *Kirkholt* Estate. The non-white populations in Rochdale are predominantly categorised as being Asian: although this category is very broad, incorporating many different ethnicities, the vast majority of this population in Rochdale are of Pakistani or Bangladeshi origin, with the second most commonly stated religious belief after Christianity being Islam.

The town of Rochdale has also been in the press because of two child sex trafficking and abuse scandals, one in 2012 (Bunyan, 2012) and a second in 2016 (BBC, 2016). The terrible nature of the crimes and the failure of the Greater Manchester Police to investigate the crimes meant the stories gained particular prominence. However, because all the criminals involved were Asian men of Pakistani–Muslim origin, and all the victims were white British females, it has been suggested the crimes were racially motivated. It has also been suggested that the Police failed to investigate because of concerns about ethnic tensions and for fear of being called racist (Bunyan, 2012). The Greater Manchester Police has since apologised for its failure to investigate the crimes (BBC, 2015).

3.3 Experimental design

We study the other–regarding behaviour of white British nationals and how their behaviour is influenced by the *English* and *Muslim* ethnic origins of those they interact with. This was achieved by conducting a door-to-door artefactual field experiment administered to the population areas outlined in Section 3.2. All subjects first completed a series of Other–Other Games, dividing £10 between two other people (Part 1). They then played a single Dictator Game, where they allocated £10 between themselves and a receiver (Part 2). All choices were made in whole pounds. Subjects then completed a post experimental questionnaire.[8]

To vary the ethnicity of the individuals that subjects are allocating money to, we provided them with their surnames. In an attempt to avoid any experimenter bias in surname selection, we classified surnames taken from the Edited Electoral Register for the Rochdale area into different groups of 'ethnic and cultural' origin using the 'Cultural, Ethnic and Linguistic' taxonomy of Mateos *et al.* (2007). Only those surnames which were classified as 'Western European, English' in origin (for example, *Smith*) and those of 'Muslim' origin (for example, *Islam*) were used.[9]

Households that answered the door were read out a fixed script that outlined who the caller was, and were asked if they would like to take part in an 'Economic Decision Making Study'. They were told they would receive £2.50 for taking part, and that they had the opportunity to earn additional money. If a resident agreed to take part the experiment was conducted at the door–step. Once finished, subjects were paid in cash. The order in which streets were approached was randomised and only one person per household was permitted to take part.

Subjects were told that once the study was completed those they were allocating money to would receive payment in cash through the post, which they did. Subjects

[8] All experimental materials are included in Appendix B.1.

[9] These names are examples, and were not necessarily used within the study. Over 400 unique surnames were employed. The surnames are not given for anonymity reasons.

were aware that they were not allocating money directly to their neighbours as the housing estates studied make up only a tiny fraction of the entire town of Rochdale. Subjects were also told that they were allocating money to people who would not be required to make a decision, and that these people were not aware that they were involved in the study. Any money they received would be a surprise. This was emphasised in an attempt to mitigate the effect that subjects' first order beliefs (their belief about the receiver's choice) and second order beliefs (their belief about the receiver's expectation of their choice) might have on their behaviour. These beliefs have been highlighted as important for reciprocity in a number of studies (Falk & Fischbacher, 2006; Yamagishi & Kiyonari, 2000). In order to control for the effects stemming from the commonality of social group affiliation (Guala *et al.*, 2013), subjects were assured they would remain anonymous to the receivers.

In Part 1 subjects were asked to allocate £10 between two other people ('*Person A*' and '*Person B*'), but were not able to allocate any money to themselves, and therefore received no payment for their decisions. They were required to do this under three schemes in which they were provided with the surnames of the two people they were allocating money to. Subjects completed what we label an *In–In*, *Out–Out* and *In–Out* scheme, in a random order. In the *In–In* scheme, subjects allocated £10 between two people with surnames of *English* origin and in the *Out–Out* scheme the money was allocated between two people with surnames of *Muslim* origin. In the *In–Out* scheme money was allocated between one person with a surname of *English* origin, and one with a surname of *Muslim* origin, the order of which was randomised between subjects to control for any order effects. Subjects were informed that one scheme would be selected for payment at random and that the two individuals from that scheme would receive payment through the post. Subjects did not learn the scheme that was selected until the experiment was completed. Part 1 is motivated by the minimal–group findings of Turner (1978), who reports that Other–Other allocation choices can enhance the salience of subjects' identities in subsequent decisions.

In Part 2 subjects were asked to allocate £10 between themselves and a receiver, an individual randomly selected from the Edited Electoral Register. The between–subject design varied whether the receiver had a surname of *English* origin (*English* Treatment), a surname of *Muslim* origin (*Muslim* Treatment) or if the surname was withheld (*Anonymous* treatment). Each subject was randomly assigned to a treatment, and the surname was unknown to the experimenter. Once they had made their choice they were paid in cash. A £10 endowment was chosen as it is the 'standard' dictator amount (Engel, 2011), and thus allows for a comparison with previous studies.

To elicit background characteristics, subjects completed a post experimental survey and self–reported a number of characteristics. However, the main survey question of interest was one which aimed to measure group attachment and to check the experimental manipulations were successful. In a manner similar to Yamagishi & Kiyonari (2000), in the post experimental survey subjects were asked, '*How close did you feel to your match in Part 2, based on their surname alone?*'.[10] Subjects were asked to make a choice on a scale from 1 to 10, with 1 being '*Not at all*' and 10 being '*Very much so*'. If ethnicity is important to subjects, then this measure of *Closeness* should register increases (or decreases) relative to the *Anonymous* baseline.[11] Table 3.2 provides a summary of the experimental design.

Treatment	*Part 1*	*No Surname*	*English*	*Muslim*	*Survey*	*Observations*
Anonymous	✓	✓			✓	38
English	✓		✓		✓	42
Muslim	✓			✓	✓	42

Table 3.2: Experimental Design – Part 2

The experiment was conducted between the hours of 12pm and 6pm during the summer of 2014. A total of 16 full days across 4 weeks were required to collect all the observations. The experiment was conducted by a single experimenter, who was a white British male. A total of 828 individual addresses from the two housing estates were approached, 341 residents answered the door, and 132 agreed to take part; 14% of the total households in the six Census areas in Table 3.1 took part. We were unable to recruit any additional residents: of the 487 addresses that did not answer the door, all were approached an additional time at a later date. The 209 addresses which refused to take part were not approached again.[12]

[10]In the *Anonymous* Treatment subjects were asked '*How close did you feel to your match?*'

[11]This question is a variation of a question used by Turner (1978), "*How much did you like the people in your group?*", or of that used by Chen & Li (2009), "*Please rate how closely attached you felt to your own group throughout the experiment*".

[12]We do not include the responses from 10 residents who were either non–white, non–British or both.

3.4 Results

This section outlines the experimental results. A number of common features are present throughout: where non-parametric tests are given, the p-value and test used are both presented in parentheses. Unless otherwise stated, the null hypothesis is always that there is no difference in behaviour between treatments and all reported tests are two sided.[13] Only one person per household took part and each individual observation is treated as independent. As the experiment was conducted over four weeks, subjects may have heard about the study from neighbours or through social media, causing behaviour to differ over time. No evidence is found of a trend in behaviour over the course of the experiment ($p = 0.76$, Cuzick's Wilcoxon–like test for trend) (Cuzick, 1985) so all observations are pooled.[14]

3.4.1 Dictator game

Table 3.3 presents summary statistics of the subjects' self reported demographics and choices.[15] Figure 3.1 displays box plots of amounts given in each treatment.

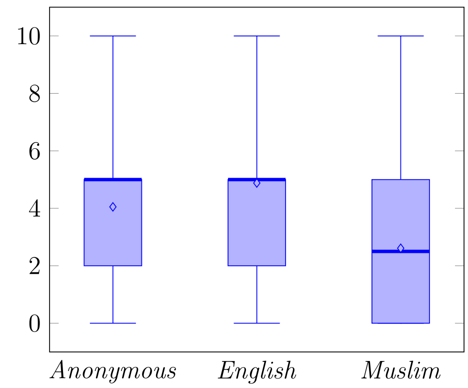
Treatment	*Male*	*Employed*	*Income*	*No Education*$^\diamond$	*Closeness*	*Amount Given*
Anonymous	57%	26%	≤£10,000	26%	2.95 (2.78)	4.05 (2.88)
English	43%	22%	≤£10,000	44%	3.78 (2.78)	4.88 (3.15)
Muslim	35%	19%	≤£10,000	29%	2.61 (2.65)	2.62 (2.47)

Note: Standard deviations in parentheses.
$^\diamond$ Percent of subjects who have no formal qualifications.

Table 3.3: Summary Statistics

Result 1 (Out–Group Negativity): *In the Dictator Game subjects give around half as much to those with surnames of Muslim origin in comparison to those with surnames of English origin, and in comparison to those who are Anonymous.*

[13] We use non-parametric Robust Rank Order tests instead of Wilcoxon-Mann-Whitney tests following the analysis of Feltovich (2003).
[14] This test is commonly used in the medical literature.
[15] See Appendix B.2 for a summary of all self reported demographics.



Note: ⋄ represents the mean.

Figure 3.1: Amounts Given in the Dictator Game, by Receivers' Surname

Support. Table 3.4 presents the test statistics and p–values from the pairwise comparisons of distributions and medians between treatments. As shown, no significant differences are reported when comparing how much was given to those with surnames of *English* origin to how much was given to those who are *Anonymous*. However, when comparing giving to those with surnames of *Muslim* origin to giving to those with surnames of *English* origin and to those who are *Anonymous*, significant differences are found in both distributions and medians.

Comparison	*Alternative Hypothesis*	*Test Statistic*	*p–value*	
Distributions⋄	H_A: *English* ≠ *Anonymous*	χ^2=1.366	0.266	
	H_A: *Muslim* ≠ *Anonymous*	χ^2=5.541	0.019	**
	H_A: *Muslim* ≠ *English*	χ^2=11.637	0.001	***
Medians+	H_A: *English* ≠ *Anonymous*	Ú=-1.115	0.265	
	H_A: *Muslim* ≠ *Anonymous*	Ú=2.341	0.019	**
	H_A: *Muslim* ≠ *English*	Ú=3.63	0.000	***

Note: In each comparison the null hypothesis is always that there is no difference between treatments. ***, **, * denotes significance at the 1%, 5% and 10% level.
⋄ Compared using Kruskall–Wallis Tests. Test statistics reported with ties.
+ Compared using Robust Rank Order Tests.

Table 3.4: Pairwise Comparisons of Dictator Giving

Table 3.5 presents estimates from a number of Tobit regressions. In each regression the amount given is the dependant variable, the *Anonymous* treatment is taken as the baseline and we include dummies for the *English* and *Muslim* treatments. To examine the robustness of the estimates, in each subsequent model additional dummies

Dependant variable:	Amount Given				
	(*i*)	(*ii*)	(*iii*)	(*iv*)	(*v*)
English$^\diamond$	1.169	1.16	0.978	0.394	-1.474
	(0.907)	(0.91)	(0.947)	(0.847)	(1.322)
Muslim$^\diamond$	-2.051**	-2.072**	-1.972**	-2.099**	-2.601**
	(0.919)	(0.938)	(0.986)	(0.885)	(1.26)
Male		-0.165	-0.533	-0.559	-0.56
		(0.757)	(0.776)	(0.689)	(0.678)
Area		0.118	0.404	0.943	0.754
		(0.753)	(0.771)	(0.691)	(0.687)
Income			0.173	0.391	0.486
			(0.72)	(0.635)	(0.627)
Employed			0.248	-0.467	-0.481
			(0.985)	(0.89)	(0.887)
High School			0.933	0.878	1.038
			(0.87)	(0.778)	(0.781)
A-Level			3.709***	3.548***	3.444***
			(1.095)	(0.975)	(0.965)
Closeness				0.732***	0.439*
				(0.156)	(0.263)
Closeness×*English*$^\diamond$					0.705*
					(0.396)
Closeness×*Muslim*$^\diamond$					0.214
					(0.371)
Constant	3.846***	3.878***	2.691**	0.998	1.698
	(0.657)	(0.847)	(1.143)	(1.107)	(1.18)
Observations	122	122	98	96	96

Note: Observations left censored at 0 and right censored at 10. Standard errors in parentheses. ***, ** and * denotes significance at the 1%, 5% and 10% level. The number of observations differs between models due to missing entries. $^\diamond$*English* and *Muslim* correspond to dummies for the *English* and *Muslim* treatments respectively.

Table 3.5: Determinants of Amount Given – Tobit Regressions

which take values of 1 for each of the following (and zero otherwise) are also included: if the subject was male (*Male*), employed (*Employed*), educated to GCSE level (*High school*), or A-level (*A-level*) and if the subject was from the *Kirkholt* estate (*Area*); when the high school dummy and A-level dummy are both zero, the subject has no formal qualifications.[16] Income (*Income*) and the subjects' self reported level of *Closeness* (*Closeness*) are also included, along with *Closeness* interacted with the treatment dummies. Supporting the non–parametric analysis, Table 3.5 outlines how the coefficient on the *Muslim* dummy is always estimated to be negative and significant ($p < 0.05$, in all regressions), with its magnitude robust to specification changes.

Although Result 1 resonates with the results of previous laboratory experiments, it does not support the typically reported result that subjects exhibit favouritism towards the in–group. Instead, Result 1 indicates how subjects exhibit *negativity* towards those perceived as an out–group.

One explanation for why we do not observe in–group favouritism could be attributed to the *English* treatment not being a strong enough experimental manipulation, or it being unsuccessful at inducing an in–group sense of identity. To shed light on this, we examine self–reported levels of group attachment, or *Closeness*, as is standard in the literature (Chen & Li, 2009; Ioannou *et al.*, 2015; Yamagishi & Kiyonari, 2000). Subjects are found to report feeling closest to those with surnames of *English* origin, with *Closeness* in the *English* treatment found to be significantly different from that in the *Anonymous* and *Muslim* treatments ($p = 0.07$ and $p = 0.011$, Robust Rank Order Tests). This is despite behaviour being found to be identical in both the *English* and *Anonymous* treatments, as outlined in Result 1. No differences in levels of *Closeness* are reported when comparing the *Muslim* treatment to the *Anonymous* treatment ($p = 0.65$, Robust Rank Order Test). This suggests the *English* treatment was successful in inducing an in–group sense of identity.

An alternative explanation may relate to the implicit and explicit attitudes of the subjects. Subjects giving to an *Anonymous* receiver may have unconsciously believed the receiver was 'like them', and therefore the same as someone with a surname of *English* origin. Only when they are explicitly prompted to consider how 'close' they

[16] Five subjects reported having higher levels of education, including further and higher education qualifications (Postgraduate and Graduate level). These, however, are not regarded as reliably self-reported and are grouped with subjects with no education. The results do not change significantly if these individuals are grouped with High School or A–level educated subjects.

feel to someone with a surname of *English* origin do they exhibit favouritism. This is likely captured by the significant, but small, effect of the interaction between *Closeness* and the *English* treatment (*Closeness*×*English*) reported in Table 3.5 ($p = 0.078$, model v). Thus, at least in this setting, in–group favouritism is only observed when explicitly prompted.

3.4.2 Other–other games

Prior to the Dictator Game all subjects played three Other–Other games where they were required to allocate money between two other people, Person A and Person B. They made their choices under three schemes in a random order: the *In–In*, *Out–Out* and *In–Out* scheme. In the *In–Out* scheme, subjects completed either an *In–Out* or *Out–In* ordering. As shown in the experimental instructions given in Appendix B.1, Person A's surname was presented first and to the left, whilst Person B's surname was presented second and to the right. We disaggregate the *In–Out* scheme data by orderings to identify any potential presentation effect.

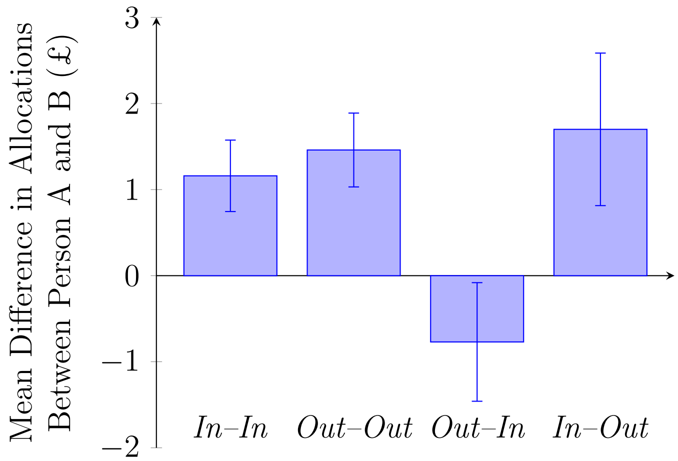
Table 3.6 presents the results from each scheme, outlining the average amounts allocated to Person A, Person B and the mean difference between these allocations. The final row of Table 3.6 presents the results from two–sided Sign Tests used to examine if these differences are significantly different from zero. The mean differences between amounts allocated to Person A and Person B for all schemes are presented graphically in Figure 3.2.

Scheme	*In–In*	*Out–Out*	*In–Out*	
Ordering			*Out–In*	*In–Out*
Person A Allocation	£5.58	£5.73	£4.62	£5.85
Person B Allocation	£4.42	£4.27	£5.39	£4.19
Mean Difference	£1.16	£1.46	-£0.77	£1.7
	(2.34)	(2.43)	(2.92)	(3.29)
Observations	122	122	69	53
H_A: Difference $\neq 0$	$p < 0.001$	$p < 0.001$	$p = 0.08$	$p < 0.001$

Note: Standard deviations in parentheses.

Table 3.6: Allocations in the Other–Other Games

Result 2 (Presentation Effect): *When the surnames are of the same ethnic origin, subjects allocate approximately £1 more to the person listed first (Person A) than to*



Note: Vertical bars represent 95% confidence intervals.

Figure 3.2: Other–Other Allocations

the person listed second (Person B).

Support. From Table 3.6, in both the *In–In* and *Out–Out* schemes, subjects give significantly more to Person A than to Person B ($p < 0.001$ in both cases, Sign Tests).

Result 3 (Discrimination): *When Person A and Person B have surnames of different ethnic origins, subjects allocate more to the individual with the surname of English origin. This is true even when the English surname is presented second. However, due to the presentation effect, we only conclusively observe discrimination in the Out–In scheme.*

Support. It can be seen from Table 3.6 that subjects give more to the person with a surname of *English* origin in the *In–Out* scheme, with the difference in both *In–Out* and *Out–In* orderings being significantly different from zero ($p = 0.001$ and $p = 0.08$, Sign Tests).

Further support for Result 2 and Result 3 is presented in Table 3.7, which outlines estimates from OLS and Tobit regressions. In each regression, the difference between amounts allocated to Person A and Person B is the dependent variable. Observations from the *In–In* scheme are taken as the baseline, and explanatory variables include dummies controlling for choices made in the *Out–Out* scheme, and the *In–Out* and *Out–In* orderings of the *In–Out* scheme, taking values of 1 in each case (and 0 otherwise).

Dependant Variable:	Differences in Allocations to Persons A and B			
	(*i*)	(*ii*)	(*iii*)	(*iv*)
Out–Out Scheme		0.295		0.3
		(0.241)		(0.242)
In–Out Ordering		0.534		0.548
		(0.488)		(0.492)
Out–In Ordering		-1.932***		-1.942***
		(0.419)		(0.421)
Constant	0.975***	1.164***	0.978***	1.164***
	(0.163)	(0.213)	(0.164)	(0.212)
Observations	366	366	366	366

Note: Observations from the *In–In* scheme are taken as the baseline. ***, ** and * denotes significance at the 1%, 5% and 10% level. Reported standard errors are clustered at the individual level. Models (*i*) and (*ii*) are OLS regressions, (*iii*) and (*iv*) are Tobit regressions.

Table 3.7: Other–Other Allocation Decisions

In support of Result 2, Table 3.7 outlines how the constant is estimated to be approximately 1 across models. The OLS regression results of model (*ii*) suggest that Person A is allocated £1 more than Person B on average. This can be inferred from the estimate of the constant. No differences are found between the *In–In* and *Out–Out* schemes, with the coefficient on the *Out–Out* dummy never significantly different from zero at conventional levels ($p > 0.1$, all cases). In support of Result 3, the coefficient on the *Out–In* ordering dummy has a negative and highly significant coefficient ($p < 0.01$). The estimates of model (*ii*) outline how, in the *Out–In* ordering, if Person B has a surname of *English* origin, once the 'Presentation Effect' is accounted for, they are allocated around £2 more than the individual with a surname of *Muslim* origin, even when that individual is presented first.

Both Results 2 and 3 are in line with previous findings in the literature. The presentation bias outlined in Result 2 is well documented in the psychology literature. In particular, there is considerable evidence of the tendency for subjects to gravitate towards options presented on the left in comparison to those presented on the right (Friedman *et al.*, 1994; Weng & Cheng, 2000). However, although Result 3 is a replication of the findings of Chen & Li (2009) and Turner (1978), and outlines how subjects differentiate between individuals conditional on their ethnicity, it is not clear how to interpret the subjects' behaviour. First, as there is no *Anonymous* baseline from which to compare behaviour, the results could equally imply either in–group favouritism or out–group negativity. Second, as behaviour in the *In–Out* scheme is indistinguishable from behaviour in the *In–In* and *Out–Out* scheme, we can only conclusively say that discrimination is observed in the *Out–In* scheme. It is possible that this is a result of a ceiling effect in how comfortable subjects feel when implementing inequitable outcomes. If subjects already hit a 'ceiling' with respect to inequitable choices in the *In–In* scheme, then it is unlikely we would then observe differences between this scheme and the *In–Out* scheme, as subjects would be unwilling to tolerate any additional inequality. Therefore, it is likely that we only observe discrimination in the *Out–In* scheme as the motive to discriminate works in the opposite direction to the presentation effect.

3.5 Structural model

To link our empirical analysis to behavioural theory, we model subjects' behaviour structurally. Following the model of Cox *et al.* (2007), it is assumed that each subject has utility

$$u(x;\alpha,\theta,\epsilon) = \begin{cases} \alpha^{-1}((s+\omega-x)^{\alpha} + \theta(e,m)x^{\alpha}) & \text{if } \alpha \in [-\infty,0) \cup (0,1] \\ (s+\omega-x)x^{\theta(e,m)} & \text{if } \alpha = 0 \end{cases}, \tag{3.1}$$

which is derived from her own payoff and the receiver's payoff. Her own payoff consists of her participation fee, $s = 2.5$, plus the amount that she keeps for herself: her initial endowment, $\omega = 10$, minus the amount given to the receiver, $x \in \{0,1,2,...,10\}$. The receiver's payoff, x, is the amount the subject decides to give. The social preference parameter, $\theta(e,m)$, is a function that captures the utility weight the subject places on the receiver's payoff. Implicit in our analysis is the assumption that subjects are making their decision in isolation, or that they are 'narrowly bracketing' their decisions, and thus, are not taking into account the annual income of the receiver or themselves (Read *et al.*, 1999). Following Chen & Li (2009), we assume this parameter is a function of the ethnicity, or identity, of both the subject and the receiver,

$$\theta(e,m) = \theta(1 + ae + bm) + \epsilon, \tag{3.2}$$

where e and m are dummy variables, with $e = 1$ when the receiver has a surname of *English* origin, and $m = 1$ when the receiver has a surname of *Muslim* origin, and 0 otherwise. Following the behavioural literature, we interpret parameter θ as capturing the utility weight placed on the payoff of the receiver in a group neutral interaction (Chen & Li, 2009). Thus, parameter θ represents baseline social preferences, or preferences when the receiver is *Anonymous*. Parameters a and b, the *English* and *Muslim* identity parameters, measure the additional effects of the receiver's *English* or *Muslim* ethnic origin on this weight. The function $\theta(e, m)$ is assumed identical across subjects, except for an idiosyncratic error term, ϵ. Thus, following the analysis of Apesteguia & Ballester (2016), we assume a random preference model rather than random utility model.

Specifying utility in this way is advantageous in comparison to more restricted forms, as the model nests many commonly assumed functional forms: when $\alpha < 1$, indifference curves are strictly convex, when $\alpha = 1$ indifference curves are linear and subjects are inequity averse (Fehr & Schmidt, 1999) and indifference curves converge to Cobb-Douglas preferences as $\alpha \to 0$. When $\theta(e, m) > 0$, as $\alpha \to \infty$ preferences are Leontief.[17] Utility reverts to standard selfish preferences when $\theta = 0$.[18] Parameters could be obtained from each of these nested forms by estimating the model with restrictions. However, we let the model pick the parameter values that best fit the data and then test to see if such restrictions would be valid. Appendix B.3 describes the strategy employed to estimate the parameters in Equation 3.1 structurally, closely following the procedure of Cox *et al.* (2007).

Table 3.8 outlines the parameter estimates and standard errors. We begin by testing a number of parameter restrictions. First, note that α is estimated to be both positive and highly significant ($p < 0.01$, Wald Test). It is found to be significantly different from one ($p < 0.01$, Wald Test), suggesting that linear, inequity averse preferences do not provide a good fit for the data. Similarly, θ is estimated to be positive and is highly significant ($p < 0.01$, Wald Test), rejecting the notion of selfish preferences.

From Table 3.8, the estimate of the *English* identity parameter, a, is not significantly different from zero at conventional levels ($p = 0.33$, Wald Test), although the *Muslim* identity parameter, b, is negative and highly significant ($p = 0.01$, Wald Test). The null hypothesis that $a = b$ can be rejected at the 5% level ($p = 0.03$, Wald Test). Social preferences towards receivers with surnames of *Muslim* origin are estimated to be around 87% smaller than social preferences towards those with surnames of *English* origin and those who are *Anonymous*. As a is not estimated to be significantly different from zero,

[17] See Cox *et al.* (2007) for the proof.

[18] If $\alpha < 0$, giving $x = 0$ would imply a payoff of $u = -\infty$. As we observe a high number of $x = 0$ in the data, we assume $\alpha \geq 0$.

Parameter	Estimate	Standard Error
θ	1.11***	0.39
a	0.65	0.67
b	-0.87***	0.36
α	0.72***	0.03
σ	2.31***	0.28
Observations	122	
Log-likelihood	-422.53	

Note: ***, ** and * denotes significance at the 1%, 5% and 10% level.

Table 3.8: Structural Parameter Estimates of Equation 3.1

subjects gain no additional utility from the payoff of individuals with a surname of *English* origin in comparison to those who are *Anonymous*. However, as b is estimated to be less than zero, subjects derive less utility from the payoffs of individuals with surnames of *Muslim* origin in comparison to those who are *Anonymous*. The estimates of a and b lend themselves to Result 1, supporting the idea that subjects exhibit out–group negativity rather than in–group favouritism, and suggest that this negativity is a consequence of group–contingent social preferences.

3.6 Counterfactual analysis

A number of important questions arise from our results: to what extent *could* our structural model explain the attitudes of the demographic of interest? Could our results provide an explanation for the reported discrimination against Muslims in both the press and other studies? For example, the audit study of Ahmed & Hammarstedt (2008) finds Swedish landlords accepting potential tenants with Swedish names more frequently than those with Muslim names, and the anecdotal evidence of Syal & Topping (2014) reports British taxi customers requesting white British drivers ("local" drivers) instead of Asian–Muslim drivers. In each case, the individual incurs some cost, or trades–off some of her income, in order to complete a transaction with someone they view as in–group, rather than someone from the out–group.

To address this question, consider an individual, p, who must choose with whom they will interact with in order to complete a transaction. They must select a strategy, s_1, from the strategy set $S_1 = \{In, Out\}$, where strategy $s_1 = In$ implies they interact with someone they perceive as an ethnically in-group player (of *English* origin), p^I, and

$s_1 = Out$ with someone they perceive as an ethnically out–group player (of *Muslim* origin), p^O. We assume that players p^I and p^O are identical except for p's perception of their ethnicity.

From strategy $s_1 = In$, p earns $\omega_{In} = \omega - c$, and from $s_1 = Out$ she earns $\omega_{Out} = \omega$, where $\omega \geq 0$ and $c \geq 0$. We assume it may be costly for p to choose an in–group interaction, having to incur a cost in order to do so. This cost, c, can be thought of as the cost of discriminating, or the amount of income individual p is willing to forego in order to pursue in–group interactions. For example, when $c > 0$, the cost could represent additional search costs associated with locating a shop run by an in–group member, or the additional wait incurred when requesting an in–group taxi driver. It could represent the cost associated with voting for a policy that causes personal cost, but would allow the individual to avoid out–group interactions. Alternatively, it could represent a landlord waiting for an in–group tenant to apply for tenancy of her properties when only an out–group tenant is currently available. Thus, c could range from very small to very large. If $s_1 = In$, p^I earns a payoff of $m_{In} > 0$ and p^O earns zero. We also assume if $s_1 = Out$, p^I earns zero and p^O earns $m_{Out} > 0$. Figure 3.3 represents the decision problem graphically.

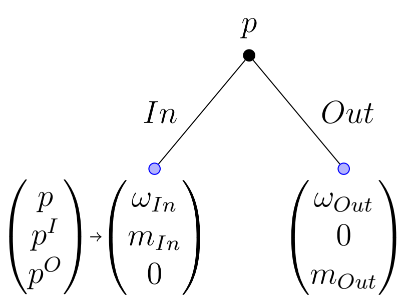


Figure 3.3: Extensive Form Representation

Assuming individual p's utility, $u : s_1 \to \mathbb{R}$, takes the form outlined in Equation 3.1, with parameters equal to those estimated in Table 3.8, $\theta = 1.11$, $\alpha = 0.72$, $b = -0.87$, $a = 0$, the extent to which she is willing to forego income in order to pursue in–group interactions can be determined. Denoting s_1^* as the utility maximising strategy choice, $s_1^* = In$ when $u(In) \geq u(Out)$,

$$\alpha^{-1}((\omega - c)^\alpha + \theta(1 + a)m_{In}^\alpha) \geq \alpha^{-1}((\omega)^\alpha + \theta(1 + b)m_{Out}^\alpha). \tag{3.3}$$

Rearranging Equation 3.3 for c as a function of m_{Out} and m_{In} gives the following two cases,

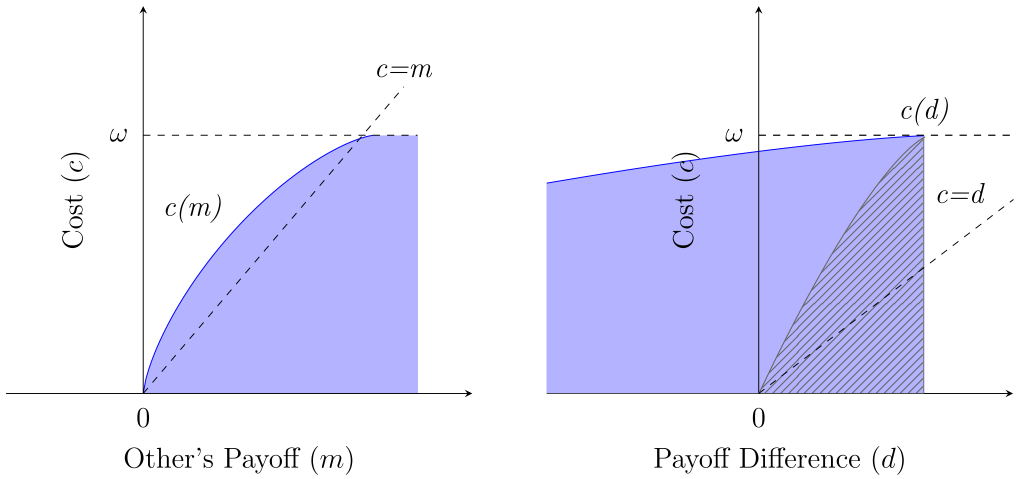
$$c \leq \begin{cases} \omega - \left[\omega^{\alpha} + \theta\left(m(b-a)\right)\right]^{\frac{1}{\alpha}} & \text{if} \quad m_{In} = m_{Out} = m \quad (3.4) \\ \omega - \left[\omega^{\alpha} + \theta\left(m_{Out}(1+b) - m_{In}(1+a)\right)\right]^{\frac{1}{\alpha}} & \text{if} \quad m_{In} \neq m_{Out} \quad (3.5) \end{cases}$$

We can now consider when p will play $s_1^* = In$ for given discriminatory costs and other players' payoffs. It is particularly interesting to consider how this choice differs between individuals with group–contingent social preferences and those without discriminatory tastes. Figure 3.4 plots Equations 3.4 and 3.5 graphically. Figure 3.4a plots costs on the y axis, and the other players' payoff m on the x acis. Figure 3.4b plots costs on the y axis, and the difference between the other players payoffs, d, on the x axis, where this difference is defined as $d = m_{In} - m_{Out}$. In each figure the light shaded areas characterise where $s_1^* = In$ is played when the identity parameters are $a = 0$ and $b = -0.87$, as estimated in Section 3.5. The dark shaded areas characterise where $s_1^* = In$ is played when the identity parameters are equal, $a = b$, or when social preferences are not group–contingent. Thus, we can consider how behaviour diverges by comparing light and dark shaded areas.

When $m_{In} = m_{Out} = m$, as shown in Figure 3.4a, p will only play $s_1^* = In$ when $a = 0$ and $b = -0.87$. In this case there are no dark shaded areas because p is always indifferent between strategies when the identity parameters are equal. However, as characterised by the areas that are not shaded, even when p has group–contingent social preferences, there still exist costs at which she would not be willing to discriminate.

When $m_{In} \neq m_{Out}$, as shown in Figure 3.4b, p may be willing to play $s_1^* = In$ even when $a = b$, as highlighted by the dark shaded area. This is due to the positive estimate of θ, which means that p would select the interaction that gave the highest payoff to the other player, regardless of their group identity. Most interesting about Figure 3.4b is that the light shaded area highlights how p would play $s_1^* = In$ for many negative values of d and even when the cost is high. This means that p would choose an in–group interaction over an out–group interaction even when that interaction would be more beneficial for an out–group player, and when the cost of selecting that in–group interaction was very high.

The conclusions drawn from this counterfactual simulation could provide an explanation for why this particular demographic may support discriminatory policies, even when those policies may produce outcomes that are costly to them, and could provide



(a) $c(m)$, if $m_{In} = m_{Out} = m$ (b) $c(d)$, if $d = m_{In} - m_{Out}$

Note: Shaded regions characterise where $s_1^* = In$

Figure 3.4: The Relationship Between Discrimination and the Other Players' Payoffs

insight into prior work that outlines discrimination towards Muslims. In particular, the analysis could provide additional insight into the results of Ahmed & Hammarstedt (2008) and the anecdotal evidence of Syal & Topping (2014). Although statistical explanations cannot be ruled out in these examples, the analysis highlights how individuals with group–contingent social preferences may be willing to incur significant costs, or be prepared to trade–off large amounts of their own income, in order to avoid out–group interactions.

3.7 Conclusion

We report evidence of individuals discriminating against those with surnames of Muslim origin and demonstrate this to be a consequence of subjects' social preferences being group–contingent. We advance the literature through the estimation of a structural model which we exploit to perform a counterfactual simulation, modelling how these preferences may cause individuals to incur costs in order to avoid out–group interactions. In doing so we provide a potential explanation for the documented attitudes of individuals from the demographic we study.

Whilst the results are suggestive, we acknowledge that care should be taken when trying to generalise the results to the behaviour of other populations and identities, and even to those who opted out of the experiment. For example, it may be that those who agreed to take part in the experiment were more inclined to discriminate than those who did not. Alternatively, as suggested by List (2006), those who participated in the experiment may be more sensitive to experimental cues unaccounted for here. However, such selection bias is unlikely to explain why we do not observe in–group favouritism. The results do, however, serve as a sign that ethnic identities embedded within surnames can have a significant effect on pro–social preferences.

Alternatively, the surnames we used could have signalled something else to the subjects which we have failed to account for (Heckman, 1998). For example, although we have focused exclusively on surnames as revealing ethnicity, Mitra & Ray (2014) outline how group conflict between Hindus and Muslims in India and Bangladesh may instead be the result of status differences. In our study population Muslims are a minority, and it may be that they are perceived as being of a lower social status. Instead of ethnicity, subjects may identify with people along these status lines. Whilst determining if this is the case is beyond the scope of this paper, the result that subjects exhibit group-contingent social preferences would still hold; the only aspect of the analysis that would change is what constitutes the social group.

In conclusion, we report evidence that the discriminatory behaviour of a previously unstudied demographic is a consequence of out–group negativity, rather than in–group favouritism. We are able to determine this because of the inclusion of an *Anonymous* baseline treatment, a crucial design aspect that is typically omitted. Our results may mean that the findings of previous studies that omit this treatment may need to be reinterpreted. In particular, the stylised conclusion that the differential treatment of individuals is a consequence of, or primarily driven by, in–group favouritism may need to be re–examined. Our results highlight how simple experiments can be used to understand the drivers of discrimination and the willingness of individuals to engage in pro–social acts.

Chapter 4

Measuring Other–Regarding Preferences in the Market for Taxis*

4.1 Introduction

Although a large number of laboratory experiments detail the prevalence and significance of other–regarding preferences, there is limited field evidence that these preferences have any implications for market outcomes (DellaVigna, 2009). Recent field studies suggest laboratory experiments may exaggerate the extent and significance of these preferences in social dilemmas (Stoop *et al.*, 2012; Winking & Mizer, 2013), possibly as a consequence of experimenter scrutiny, the decision context, self–selection of participants, stake sizes, or the artificial restriction of choice sets that the lab imposes (Levitt & List, 2007). Other studies highlight the importance of reputational concerns (List, 2006) and monitoring considerations (Bandiera *et al.*, 2005; Benz & Meier, 2008) in explaining what might otherwise be considered as other–regard in natural settings. These criticisms and concerns raise serious questions about both the generalisability and interpretability of laboratory experiments that measure other–regarding preferences, and the importance of these preferences for economic outcomes.

Other–regarding preferences also form the foundation for recent behavioural theories of discrimination. Stemming from concepts of 'taste–based' discrimination first detailed in Becker (1971), a prominent theory is that social preferences are group–contingent, or that other–regarding preferences are larger towards those we identify with (the 'in–group'), in comparison to 'out–groups' (Chen & Li, 2009). Although this

*The field experiment reported in this chapter was conducted in collaboration with Brit Grosskopf.

explanation has gained prominence, as with other work on social preferences, the majority of evidence in its support has been obtained from laboratory experiments (Chen & Chen, 2011; Drouvelis & Nosenzo, 2013; Goette *et al.*, 2006; van Der Mewe & Burns, 2008). Field experiments, in contrast, largely suggest discriminatory behaviour can be attributed to statistical discrimination (List, 2004; Levitt, 2004; Gneezy *et al.*, 2012), although some come close to identifying a taste (Bertrand & Mullainathan, 2004; Mujcic & Frijters, 2013). In addition, the methods used for studying identity and discrimination in the laboratory have recently been criticised, with work suggesting the observed behaviour is a consequence of experimenter demand effects (Zizzo, 2010, 2012), or possibly stemming from a heuristic (Guala & Filippin, 2015), rather than being due to an inherent preference.

The purpose of this paper is to examine the prevalence, and extent, of other–regarding preferences in transactions that take place in a highly competitive market setting, and determine the role played by reputational concerns in fostering other–regarding behaviour. We also investigate the significance of ethnic identity in determining these preferences, and examine its interplay with individuals' reputational concerns. This is done using a natural field experiment whereby we employed 22 testers of varying ethnicity to pose as passengers and take a number of pre–determined taxi journeys.[1] In each case we endowed them with only 80% of the expected fare. Once the taxi meter reached 60% of the fare, testers told the driver that they only had a certain amount, and asked if they could have the final 20% of the journey for free. The tradeoff faced by a driver in this situation is analogous to the dilemmas that subjects typically face in the laboratory: express other–regard at a personal cost but to the benefit of another by giving some of the journey for free, or to behave selfishly but profitably by stopping once the meter reaches the amount the passenger can afford. Although this transactions is not competitive *per se*, as the passenger has no alternative driver with which to interact, the taxi market setting we study satisfies all the requirements of a market place, as discussed by Al-Ubaydli & List (2016). As such, the driver makes his decision within a competitive context, and where the objective function is typically assumed to be to maximise profits.

In a 2×2 between–subjects design we systematically vary the length of the taxi journeys using *Short* and *Long* distance treatments, where testers took journeys of approximately 1.7 miles and 4.4 miles. As drivers assigned to the *Long* distance treat-

[1] Under the taxonomy of Harrison & List (2004) our experiment is classified as a natural field experiment.

ment are able to give twice as much (in absolute terms) as drivers assigned to the *Short* distance treatment, we can examine if the drivers' other–regarding preferences depend on the relative payoffs between themselves and the passenger, or if giving is constant regardless of the amount available to give. Orthogonally to the distance treatments, we vary whether the drivers' reputations are a concern to them or not. Using a *No Reputation* treatment, testers signal the one–shot nature of the interaction to the driver. The taxi markets we study have thousands of drivers, and tens of thousands of passengers each week, making repeated interactions for infrequent customers incredibly unlikely; these markets are therefore attractive for studying the 'one–shot' interactions required for disentangling other–regard from reputational concerns. As described in Section 4.3, the only real possibility of meeting a driver in a future interaction is by obtaining his contact details so that he can be actively selected. Our *Reputation* treatment, similar to the repeat business treatment of Schneider (2012), exploits this with testers asking drivers for a business card so they can contact them for future journeys. Making drivers' reputations salient will allow us to examine how the prospect of a repeated interaction affects drivers' other–regarding behaviour.

We find that 70% of drivers in the *No Reputation* treatment give part of the journey for free, with more than 25% completing the journey at no extra cost to the tester. We also find that the extent of giving is proportional to the length of the journey. Drivers give around 10% of the expected fare in both *Short* and *Long* distance treatments. In the *Reputation* treatment, we observe only 45% of drivers giving out a business card when asked, and although giving is increased slightly on average, reputational concerns have no significant effect on their other–regarding behaviour.

Differential treatment of testers, conditional on both their own and the drivers' ethnicity, is also observed: white and South–Asian drivers give significantly less, and are significantly less likely to complete a journey when the tester is black. This result is robust to a comprehensive range of field, journey, driver and tester specific variations obtained from each individual journey. Tester specific characteristics are obtained from a complementary laboratory experiment, following the procedure of Xiao & Houser (2005). We elicit the perceived aggressiveness, attractiveness, friendliness, trustworthiness and wealthiness of the testers' appearance, traits that are otherwise 'unobservable', but may vary with ethnicity (Heckman, 1998). To link our results to behaviourial theory, we also conduct a structural analysis in order to obtain other–regarding preference parameter estimates. Estimates from a range of models reveal that the other–regarding preferences of drivers are qualitatively and quantitatively similar to those obtained from laboratory experiments, and that these preferences are group–contingent.

In the *Reputation* treatment we find that reputational concerns can increase the drivers' other–regard, but only when drivers are carrying a white tester. Black testers see no significant increases, and we observe decreases in giving for South–Asian testers stemming from reputational concerns. The differential effect of reputation is attributed to the drivers' beliefs being influenced by the passengers' ethnicity, either their belief about the probability of a repeated interaction, or their belief regarding the payoff they will receive from the future interaction. This is discussed further in Section 4.5.

This study makes a number of contributions. First, we contribute to the debate on the generalisability of laboratory experiments by providing evidence that other–regarding preferences can appear in transactions observed a natural competitive market setting with a similar prominence to that observed in the laboratory. Our findings are in contrast to the evidence from the field study of List (2006), but also that of Stoop *et al.* (2012) and Winking & Mizer (2013), although in line with the findings of Stoop (2014). Second, we find evidence that the effects of reputational concerns on behaviour are not as strong as theory might predict. Finally, we find evidence to suggest that discrimination can manifest itself within beliefs as well as other–regarding preferences, in line with recent behavioural theories (Chen & Li, 2009; Chen & Chen, 2011).

The remainder of this paper is organised as follows. Section 4.2 reviews the relevant literature, Section 4.3 discusses the taxi markets we study and Section 4.4 outlines the experimental design in detail. Section 4.5 outlines reduced form estimation results, and estimates from a structural model. Section 4.6 examines the robustness of our results by accounting for potential multiple hypothesis testing. Section 4.7 discusses alternative interpretations of the results and Section 4.8 concludes.

4.2 Literature

4.2.1 Other–regarding preferences

As highlighted in the reviews of Camerer & Fehr (2004) and Cooper & Kagel (2009), other–regarding preferences are well established to exist in the laboratory. However, it is typically assumed that social preferences are irrelevant in market settings (Schmidt, 2011) and, as reported by DellaVigna (2009), there is little field evidence to support many of the laboratory derived conclusions. Levitt & List (2007) provide a range of reasons for why these conclusions may fail to generalise to field settings. In addition, pro–social behaviour in the field is often difficult to attribute to inherent preferences, as it is easily attributed to reputational concerns and social pressure effects (Akerlof & Kranton, 2000).[2]

Laboratory experiments have typically focused on dictator games, ultimatum games and public goods games in order to measure social and other–regarding preferences. However, individuals determined to have such preferences are often observed to behave selfishly under different institutions; competitive settings appear to 'crowd out' other–regarding behaviour. For example, it is well established that individuals reject unfair offers in ultimatum games, suggesting that subjects are inequality averse (Fehr & Schmidt, 1999). Yet, many experimental markets converge on the competitive equilibrium.[3] Whilst some suggest this result is indicative that individuals do not have these preferences, the models of Fehr & Schmidt (1999) and Bolton & Ockenfels (2000) predict this outcome. This behaviour could be explained by individuals being unable to enforce an equitable outcome within a market setting, and so they make the best of a bad situation (Camerer & Fehr, 2006). Dufwenberg *et al.* (2011) show this theoretically in a general equilibrium framework: under certain conditions, the market behaviour of agents with other–regarding preferences cannot be distinguished from those with standard preferences.[4]

A serious criticism raised against measuring other–regarding preferences in the lab is the influence of experimenter scrutiny on the behaviour of subjects (Levitt & List, 2007). As highlighted by Zizzo (2010), the obtrusiveness of the laboratory may encourage subjects to behave how they believe the experimenter wants them to, or how they should, rather than how they would otherwise. In support of this argument, Hoffman *et al.* (1996) show how increasing the level of anonymity granted to subjects, by moving to a 'double blind' procedure in the dictator game, drastically reduces the amount of giving. Haley & Fessler (2005) find that a pair of eyes on the screen drastically increases giving. Similar findings have been reported in bargaining games (Hoffman *et al.*, 1994). From the field, Winking & Mizer (2013) analyse the dictator game in a natural field experiment in Las Vegas, giving strangers at a bus stop $10 worth of casino chips, and suggesting they share them with another stranger. When the stranger is aware they are being scrutinised by an experimenter, they behave in line with the laboratory

[2]There is a rich literature examining the robustness of reciprocity and gift–exchange in the field, e.g. Gneezy & List (2006) and Falk (2007).

[3]See Roth *et al.* (1991) as an example.

[4]Schmidt (2011) provides an excellent review of this literature.

predictions, but when they are unaware they behave perfectly selfishly. Scrutiny appears to encourage pro–social behaviour.

However, there is considerable evidence against this criticism. For example, in contrast to Hoffman *et al.* (1996), Koch & Normann (2008) do not observe decreased giving as the level of anonymity is increased and Bolton *et al.* (1998) cast doubt on the bargainning results of Hoffman *et al.* (1994). In addition, there is increasing evidence of correlations between laboratory and field behaviour. Benz & Meier (2008) analyse how charitable giving behaviour correlates between the lab and the field, and find reasonably high correlations between behaviours two years apart. In contrast to Winking & Mizer (2013), Stoop (2014) finds strong evidence of dictator giving in a natural context, and finds that varying the level of scrutiny the subjects are under has no effect on giving rates. Our experiment adds to this literature by measuring other–regarding preferences with no experimenter or third–party scrutiny.

There is also evidence that the decision variable through which individuals express pro–social behaviours can influence their decisions. A seminal study by Stoop *et al.* (2012) studies the behaviour of fishermen in a social dilemma game. They build a bridge between the laboratory and the field, first analysing behaviour in the laboratory in a standard VCM (voluntary contribution mechanism) game, then behaviour at the bank of a fishing pond in the same game, and finally in a framed field experiment where they induce a VCM game through actual fishing. Although the fisherman behave highly other–regarding in the laboratory and at the bank of the pond, once the task is changed to fishing, no cooperation is observed. A real task *reduces* cooperation in comparison to a virtual one. Our study makes a similar contribution, as we analyse behaviour from a real task directly associated with a particular job.

In the field, DellaVigna *et al.* (2012) use a novel natural field experiment, nested within a charitable door–to–door fund raiser, in order to disentangle altruism from social pressure effects. The experiment gives potential donors the option to opt–out of meeting a fund raiser, allowing those who might give as a consequence of social pressure to select out. Although they find that a significant number of individuals give out of pure preference to do so, social pressure is found to increase giving substantially. There is also evidence to suggest social pressure influences voter turn–out (Gerber *et al.*, 2008), and causes workers to partially internalise the negative externalities of free riding; Mas & Moretti (2009) find that worker effort is positively related to the productivity of workers who observe them, but also with those they expect to interact with again.

List (2006) considers the behaviour of local and non–local sports card dealers, where the former have reputational concerns whilst the latter do not. List finds that the locals exhibit gift–exchange, but the non–locals do not, interpreting this as gift–exchange driven by reputational concerns, although alternative interpretations of the data have been proposed by Camerer (2015), and subsequently critiqued by Al-Ubaydli & List (2015). Other studies find reputational concerns to have a minimal impact on behaviour. In a field experiment, Schneider (2012) finds that car mechanics are only influenced by the prospect of repeated interactions in certain transactions. An important conclusion, is that the predicted effect of reputation on behaviour depends heavily on the assumptions of the model being used to predict the outcome. In the lab, Grosskopf & Sarin (2010) find that when reputational concerns and social preferences are at odds (Ely & Valimaki, 2003), the latter is likely to surpass the former. The authors report strong evidence that, even when faced with reputational concerns, individuals take others' interests into account. As a result, they show that the effects of reputation on behaviour are not as large as theory predicts. They further provide evidence against the implicit argument of List (2006): that reputation and social preferences are substitutes.

4.2.2 Discrimination

Within economics, both laboratory and field experiments have been used to examine the role that ethnic and gender identities play in shaping behaviour. Laboratory experiments can be divided into those studying natural identities, such as race, gender and ethnicity, and those examining induced group identities using variations of the minimal group paradigm (Tajfel *et al.*, 1971). Framed field experiments are similar in design to laboratory studies, whilst natural field experiments are typically either audit or correspondence studies.

Laboratory studies have considered the implications of natural identities for behaviour in a number of social dilemmas. In dictator games ethnicity (Whitt & Wilson, 2007), race (van Der Mewe & Burns, 2008), political views (Fowler & Kam, 2007) and Jewish identity (Fershtman & Gneezy, 2001) have all been shown to produce favouritism towards a particular social group. In prisoners' dilemma games, cooperation rates are increased when kibbutz members play with each other (Ruffle & Sosis, 2006) and when members of the same randomly assigned platoon play together (Goette *et al.*, 2006). Further, members of minority ethnic groups display greater cooperation rates towards each other than those of ethnic majorities do towards each other (Cox *et al.*, 1991). However, it is often unclear *why* these identities affect behaviour in these ways. This is largely due to the complex, and often ambiguous ways in which identities interact, making it difficult to distinguish between *taste based* discrimination (Becker, 1971) and *statistical* discrimination.

In order to try and understand discriminatory behaviour in the absence of stereotypes and beliefs that may otherwise affect behaviour, laboratory experimenters have turned to study artificially induced identities by using minimal (Tajfel *et al.*, 1971), near–minimal and enhanced group paradigms (Chen & Chen, 2011). Through inducing an artificial identity in the lab, the experimenter can control the identity that guides behaviour, thus removing the ambiguities and complexities that arise from studying natural identities. Evidence from these paradigms suggests that other–regarding and pro–social behaviours are larger when individuals interact with those they identify with (the 'in–group'). They behave more charitably in dictator games and more reciprocally in trust games (Chen & Li, 2009). Common identities result in leaders contributing more in sequential public goods games (Drouvelis & Nosenzo, 2013), and pairs to coordinate more efficiently in minimum effort games (Chen & Chen, 2011). The prevailing explanation for these effects, which has recently been criticised (Zizzo, 2012; Guala & Filippin, 2015), is that social preferences are group–contingent (Chen & Li, 2009), or that an individual's other–regard is conditional on how they identify with the person they are interacting with.

A common type of field experiment designed to analyse discrimination in labour markets are correspondence studies, in which the experimenter fabricates a large number of identical CVs whilst varying either the ethnicity, nationality or gender of the applicant through the use of names, or photos. In a seminal study into discrimination conducted in the US, Bertrand & Mullainathan (2004) examine the extent to which employers treat applications with stereotypically black names differently to those with stereotypically white names in job call back decisions. Applications with white names receive 50% more call backs than those with black names. Similar findings have been reported in Australia, across multiple minority ethnic groups (Booth *et al.*, 2012), and in Canada across multiple occupations (Oreopoulos, 2011). Such studies come close to identifying a 'taste' for discrimination, although statistical discrimination can often not be ruled out.

Those studies which are most related to ours, audit studies, utilise actors to take part in standardised interactions such as job interviews.[5] These studies have typically

[5] See Riach & Rich (2002) for a survey of audit studies.

used 'pairs' of people matched on observable characteristics, with the implicit assumption that they differ *only* by, for example, ethnicity or gender. The most prominent audit studies report evidence of statistical discrimination. List (2004) finds evidence that sports card sellers charge buyers from minority ethnic groups more for the same card than white buyers. However, this is attributed to those minority buyers having higher reservation values, rather than being the result of taste based discrimination. Gneezy *et al.* (2012) conduct a series of experiments designed to parse taste based and statistical discrimination. Although the majority of evidence points towards statistical discrimination, weak evidence in favour of the taste based explanation is found in the treatment of homosexuals. They conclude that further study is required.

A number of audit studies of taxis report statistical discrimination by drivers, along both ethnic and gender lines. Castillo *et al.* (2013) find evidence that male taxi drivers in Peru discriminate in favour of women by agreeing to lower fares when bargaining over identical journeys. Similar to the findings of List (2004), this is attributed to men having higher reservation values than women. Further evidence from Balafoutas *et al.* (2013) suggests that drivers in Athens, Greece, take non–locals on a longer, and therefore more expensive route, than locals for journeys to the same destination. Although this appears to be the result of taste based discrimination against foreigners, such behaviour is consistent with drivers exploiting informational asymmetries between passengers, as non–locals are unlikely to be familiar with the average fare of a particular journey. Using observational data, Jackson & Schneider (2011) detail how New York City taxi drivers who lease a car from a member of their country–of–birth exhibit reduced effects of moral hazard. They argue that such a result is consistent with the presence of increased social sanctions in the form of community–enforced punishments, rather than being a consequence of social preferences.

Whilst not a study of taxis, the study closest to ours is that of Mujcic & Frijters (2013). Exploiting a natural interaction between bus drivers and passengers, paid testers acting as passengers attempted to board buses without any money. They find that white testers are allowed to embark 72% of the time, Indians 51% and blacks just 36% of the time. This result remains robust to a wide range of controls, including tester characteristics, such as aggression, attractiveness, and others, elicited through a post–experimental survey. These controls, which are neither elicited in an incentive compatible nor in an anonymous manner, are included in an attempt to control for the 'Heckman criticism' (Heckman, 1998): implicit in the assumptions of all audit studies is that *unobservable* characteristics of confederates are identical across gender or ethnicity.

The interaction can be viewed as an other–other allocation game (Tajfel *et al.*, 1971; Turner, 1978), where the driver must allocate resources between the passenger and the bus company, rather than being comparable to the dictator game. As drivers are not monitored, their choices, while costly to the bus company, are financially costless to them. Our study distinguishes itself from Mujcic & Frijters (2013) as we consider discrimination in a situation where pro–social behaviour is costly to the person exhibiting it.

As highlighted by Heckman (1998), a common misconception is that tastes for discrimination will disappear from markets in the long run. However, this is only the case under certain market conditions. The example that Heckman gives is of entrepreneurs and their hiring decisions: if entrepreneurs have a taste for white employees over those that are black, they can indulge this taste as long as they gain income. Only if the supply of entrepreneurship is perfectly elastic in the long run *at a zero price*, so that entrepreneurs have no income with which to indulge their tastes, will taste based discrimination disappear.

4.3 The market for taxi services

In the United Kingdom, there are two types of vehicles that operate as taxis: private hire vehicles (PHVs) and Hackney carriages. PHVs are not as strictly regulated as the latter, and anyone who has a driving license and is willing to pay the licensing fee, in practice, is able to become a PHV driver. PHVs are unable to ply for hire and must be pre-booked over the phone: passengers must actively select a company or driver for a given journey. The price of the journey (or fare) is independently set by each firm, or negotiated ex-ante, and vehicles often don't have a fitted meter. As such, PHV fares can vary wildly, as can the types of vehicles used.

In contrast, Hackney carriages are taxis in the true sense: drivers can ply for hire, with customers able to hail or call them, and drivers are able to wait at designated taxi ranks to be approached by customers. Drivers and passengers are randomly matched, and importantly, customers are unable to select their driver. When hailing a vehicle, a customer must take whichever driver happens to be in the area. At a rank, customers must take the taxi at the front of the queue, and drivers further down the queue will refuse journeys from customers who approach them. The only real possibility of using the same driver repeatedly is by obtaining his personal contact details.

	Birmingham	Greater Manchester: Manchester	Greater Manchester: Trafford	Greater Manchester: Salford
Initial Charge	£2.20 (187 yards)	£2.30 (404 yards)	£2.00 (815 yards)	£2.40 (480 yards)
Mileage Charge	20p per 125 yards to 1062 yards	20p per 190 yards	20p per 164 yards	22p per 240 yards
Thereafter	20p per 195 yards	-	-	-
Wait Time Charge	20p per 45 seconds	20p per 39 seconds	28p per 60 seconds	20p per 90 seconds
Cost of 1.8 Mile Journey	£5.44	£5.26	£4.87	£4.86
Cost of 4.4 Mile Journey	£10.12	£10.16	£10.61	£9.17
Wait Time Rate (per hour)	£16.00	£18.46	£16.80	£8.00

Source: Fare information is taken from Birmingham, Manchester, Trafford and Salford Council 2015 taxi fare tables obtained through correspondence with the respective licensing authorities. Manchester, Trafford and Salford are boroughs within the Greater Manchester area. All calculations based on a journey made by a single passenger with no luggage, between 9am and 5pm.

Table 4.1: Taxi Fares by Local Authority

The strict regulation of Hackney carriages ensures their similarity, with all drivers having to pass a road knowledge and English language test. All vehicles have to adhere to strict standards, such as being fitted with safety screens to separate the driver and passenger, having wheel chair access and the vehicle being under a certain age.[6] All vehicles are fitted with a taxi meter which displays the cost of the journey, up to a given point, to the passenger. The meter starts from a fixed amount and increases by a set amount every so many yards driven, or seconds waiting in traffic. Metered fares are set by the local authority. Those relevant for this study are detailed in Table 4.1.

Important to our study is the fact that the metered fare is the maximum fare the driver is able to charge the passenger unless a different fare was negotiated prior to the passenger entering the taxi. If no ex-ante negotiation took place, the metered fare is the amount the passenger must pay by law. Where no negotiation took place, fare reductions are made entirely at the driver's discretion and the driver is within his rights to refuse any reductions the passenger asks for. The 2014 *Birmingham Unmet Taxi Demand Survey* indicates that the vast majority of Hackney carriages (90%) are

[6] This is the case in the cities that we study, but varies throughout the UK.

		Greater Manchester	
Local Authority	Birmingham	Manchester	Trafford
Number of Taxis	1,255	1,086	143
Number of Ranks	19	49	18
Top five taxi ranks, ordered by weekly passenger numbers:			
1	13,611	19,109	2,447
2	4,102	5,953	2,309
3	2,686	4,312	1,743
4	2,457	3,750	833
5	2,093	3,189	530
Total Per Week:	45,778	56,830	9,033

Source: The number of operating Hackney carriages is taken from the Birmingham (2014), Manchester (2012) and Trafford (2015) *Unmet Taxi Demand Surveys* and from correspondence with the licensing authorities of the respective councils. No information was made available by Salford Council, except that there are 111 operating taxis. The figures presented here exclude hailed and pre–booked journeys.

Table 4.2: Taxis, Taxi Ranks and Weekly Passenger Numbers

driver owned: drivers keep all the fare, any tips (which are typically around 10%), and incur all the costs associated with a journey.[7] The cost of a discretionary fare reduction is therefore borne exclusively by the driver.

The markets we study are incredibly thick, with tens of thousands of journeys taken each week, with over a thousand licensed Hackney carriages operating in each city. As outlined in Table 4.2, some of the taxi ranks see over 19,000 passengers per week. The sheer number of transactions, large number of taxi ranks and the ability of drivers to 'cruise' streets plying for hire, means an infrequent user of Hackney carriages is highly unlikely to have a repeated interaction with the same driver, and the driver they do interact with is essentially randomly assigned.

4.4 Experimental design and procedure

The experiment was designed to measure other-regarding preferences of Hackney carriage drivers (herein taxi drivers) in actual market transactions, and determine the extent to which these preferences vary with their own and the passenger's ethnicity. It was also designed to examine if reputational concerns can explain other-regarding

[7] Many drivers are, however, affiliated with a firm from which they can take private hire bookings.

behaviour. We use a natural field experiment that allows us to observe behaviour in a market setting, in a natural interaction devoid of experimenter scrutiny. Our subjects, the taxi drivers, were oblivious to a study taking place.

4.4.1 Testers

The testers were hired by placing a job advert looking for 'Research Assistants' on the *Universal Jobsmatch* website, a national website initiated by the UK government's Department for Work and Pensions which anyone can use to advertise a job. The advert stated that individuals were required to assist in conducting some 'economic research'. Although the specific job role wasn't stated, it was advertised that some walking in and around the city centre would be required. Everyone who applied was invited to attend a briefing and training session at a neutral location, where they were told about the job role and asked to sign consent forms in order to take part. The rate of pay was £8.30 per hour (all experimental materials are given in Appendix C.1).

Briefing sessions lasted between 1 and 2 hours and a single treatment was discussed in detail. Testers were given copies of *one* script they were required to follow, and the experimental sheet they would have to complete.[8] They were told the script may vary, and that they would be given a chance to practice any variants before completing the task. Testers were told explicitly to follow the script as closely as possible, and when interacting with the drivers they were told they must not attempt to influence any of their decisions. Testers were told not to engage in conversation with the drivers, and scripted responses were given to anticipated questions. Our hypotheses and predictions regarding the study were never made clear to the testers, and not all the testers met each other, reducing the opportunity for testers to guess the study might involve their own ethnicity.[9] All testers wore casual clothing.

Each tester also consented to have their face photographed for 'research purposes'. Once the experiment was complete, we had their appearance rated by subjects in a follow–up laboratory experiment. Subjects in the lab had to rate the pictures for aggressiveness, attractiveness, friendliness, trustworthiness and wealthiness, on a scale from 1 to 10 (with 1 being '*Not very*' and 10 being '*Very*'). This was done to control for otherwise unobservable characteristics that may vary with the testers' ethnicity

[8]We discussed the *Short* distance / *No Reputation* treatment, which is described in Section 4.4.2.

[9]Once the study was completed, all the testers were asked to guess what they thought the study was about. None correctly identified the research questions.

(Heckman, 1998). These 5 characteristics were chosen for a number of reasons. First, the importance of an individual's attractiveness in fostering the helping behaviours of others has been outlined in a wealth of studies, with the most attractive typically found to be treated most generously (Benson *et al.*, 1976). Attractiveness has also been shown to be successful in promoting others' other-regarding behaviours (Landry *et al.*, 2006) and is correlated with labour market outcomes (Mobius & Rosenblat, 2006). Secondly, historical and recent evidence suggests that faces that appear aggressive and unfriendly, or threatening, may stimulate a different thought system in comparison to one seen as non-threatening. For example, Öhman (1986) argues that threatening faces activate the 'fear system' and therefore provide a powerful stimuli. If this is the case, faces displaying differing levels of aggression and friendliness may trigger different types of behaviours, such as self-defensive compared to helping behaviours (see Schupp *et al.* (2004) for evidence, and a discussion of the literature). Thirdly, any differential in giving stemming from ethnicity may be related to status differences relating to wealth, similar to that shown by Mitra & Ray (2014). Finally, as the interaction between a driver and tester may rely on the driver trusting the passenger regarding how much money they have, we also elicit the passengers' facial appearance of trustworthiness.

To obtain the ratings, each laboratory subject was shown a random set of 11 photos and asked to rate their appearance. Following Xiao & Houser (2005), to increase subjects' attentiveness to the task they were told that one photo, and one characteristic of that photo, would be selected at random, and if their decision for that photo and that characteristics was in line with the ratings of the majority of the other subjects in the session, they would receive £2. It took subjects around 10 minutes to rate all the photos required of them. A sample of 1188 ratings was obtained from 108 laboratory subjects. The ratings are presented in Table 4.3.[10,11]

We find that black testers are rated significantly less attractive, trustworthy, friendly and wealthy than both white and South–Asian testers ($p < 0.001$ in all cases, Robust Rank Order Tests). Black testers are also rated the most aggressive ($p < 0.001$ in both cases, Robust Rank Order Tests). Interestingly, white testers are rated as less attractive, trustworthy, friendly and wealthy than South–Asian testers ($p = 0.06$ for attractiveness, $p < 0.001$ in all other cases, Robust Rank Order Tests). White testers are also seen as more aggressive than the South–Asian testers ($p < 0.001$, Robust Rank

[10] Table C.1, in Appendix C.2, presents the correlations between the Testers' perceived facial appearance characteristics.

[11] The photo ratings sessions were conducted at the end of other, unrelated experimental sessions.

	All testers	Tester Ethnicity		
		White	Black	S.–Asian
Age	27.6	29.5	26.14	24
	(8.25)	(10.18)	(5.64)	(4.58)
Gender (1 if *male*)	0.68	0.58	0.86	0.67
	(0.477)	(0.52)	(0.378)	(0.58)
Aggressiveness	4.02	3.98	4.61	2.86
	(2.28)	(2.30)	(2.32)	(1.61)
Attractiveness	4.73	4.81	4.43	5.16
	(2.08)	(2.15)	(2.01)	(1.86)
Friendliness	5.92	5.86	5.52	7.07
	(2.25)	(2.24)	(2.27)	(1.85)
Trustworthiness	5.68	5.69	5.19	6.76
	(2.15)	(2.13)	(2.17)	(1.74)
Wealthiness◇	5.27	5.46	4.56	6.21
	(1.85)	(1.90)	(1.65)	(1.46)
No. of Ratings	1188	638	383	167
No. of Testers	22	12	7	3

Note: Testers' age and ethnicity is self-reported. Correlations between appearance characteristics are presented in Table C.1 in Appendix C.2. The raters' ethnicities are presented in Figure C.1 in Appendix C.2.
◇ Wealthiness ratings were obtained from 60 laboratory subjects, with the following total ratings: 660 across all testers, 360 for white, 210 for black, and 90 for South–Asian testers.

Table 4.3: Tester Characteristics

Order Test). We control for these tester specific variations in our parametric analysis in Section 4.5.

We focus on facial appearance due to the way that the driver and tester interact whilst in the taxi. As outlined in Section 4.4.2, the driver's decision to behave other–regarding is made whilst he is driving, and so he is likely to view the tester briefly, either through his rear-view mirror, or by looking over his shoulder. Visual emphasis will be placed on the tester's face, rather than other physical traits such as their BMI, height or build.

4.4.2 Procedure

On a given day, a tester was blindly and randomly assigned to a treatment and was required to complete between 3 to 10 journeys. As the journeys were taken from ranks, the tester had to approach the taxi at the front of the rank, enter the taxi and then

		Short Distance	*Long Distance*
No Reputation	Entry Script	*"I don't take taxis very often."*	
	Endowment	£4	£8
	Expected Fare	£5	£10
Reputation	Entry Script	*"I'm looking for a reliable driver for future journeys. Can I have a business card?"*	
	Endowment	£4	£8
	Expected Fare	£5	£10

Note: The expected fare of journeys in each treatment is approximate.

Table 4.4: Experimental Design Summary

state their destination. The experiment first varies the distance of the journeys in *Short* and *Long* distance treatments, with journey lengths of approximately 1.7 miles and 4.4 miles, which had expected fares of approximately £5 and £10. The testers were endowed with either £4 or £8 for each journey, depending on its distance. Journeys were taken in either Birmingham or the Greater Manchester area, with those starting in Birmingham taken over 5 days, and those in Manchester over 3. All journeys were taken between 11am and 5pm and at least 4 testers were in the field at any given time, along with an experimenter.

Upon entering the taxi, the tester first stated their destination, and then spoke a simple entry statement.[12] In the *No Reputation* treatment they stated, "I don't take taxis very often", and in the *Reputation* treatment they stated, "I'm looking for a reliable driver for future journeys. Can I have a business card?". The first statement signals to the driver that the interaction is one-shot, as a passenger who doesn't take taxis very often is unlikely to meet the same driver twice. The second statement is designed to signal that a repeated interaction is possible, that the drivers' behaviour may influence the probability of a future interaction, and may affect the payoffs from a future interaction.[13] The scripts were designed to be kept simple in order to keep them standardised and to avoid actor bias (Heckman, 1998), but also to keep them natural and believable to the drivers. This design feature clearly contrasts with laboratory experiments, where interactions are designed to be 'sterile' and, predominantly, without context.

Once the taxi journey began, the testers were required to wait in silence until the

[12]The first ride taken by each tester was discreetly observed by the experimenter, to ensure they entered the taxi correctly.

[13]For example, by affecting the amount the passenger tips.

meter reached a certain amount: £3 in *Short*, and £6 in *Long* distance journeys, or 60% of the expected fare. Once the meter reached this amount, testers spoke the following endowment statement: "I'm sorry, I only have £x! Can you still take me to my destination for that amount?", where $x =$ £4 in *Short*, and $x =$ £8 in *Long* distance journeys. By revealing this to the driver once the meter reached 60% of the expected fare, the driver was given ample time to stop the taxi. It also signalled the testers' intention to pay the amount that they could afford, removing any belief the driver may have that the passenger won't pay. Table 4.4 summarises the experimental design.

Driver Characteristics					
		Driver Ethnicity			
	All Drivers	White	Black	South–Asian	Other
Age	44.3	50.06	40.36	42.6	41.3
	(10.67)	(10.56)	(9.36)	(10.03)	(11.45)
Gender (1 if *male*)	0.99	0.97	1	0.99	1
	(0.12)	(0.17)	(0)	(0.07)	(0)
Journeys	283	71	11	191	10

Field Characteristics	Mean
Traffic (1 if Not Busy, 10 Very Busy)	4.44
	(2.26)
Weather (1 if *raining*)	0.11
	(0.32)

Ride Characteristics	Mean
Conversation (1 if driver attempted a *conversation*)	0.28
	(0.45)
Cashpoint (1 if driver offered a *cashpoint*)	0.04
	(0.2)
Business card, *Reputation* only (1 if one was given)	0.45
	(0.5)
Receipt Given (1 if given)	0.9
	(0.308)

Note: Standard deviations in parentheses. Where the driver's ethnicity is classified as 'Other', the tester either did not complete the experimental sheet, or classified them outside the 3 main ethnic groups that are specified.

Table 4.5: Variables Recorded by the Testers

We refer to the driver continuing the journey past the amount that the tester can afford as *giving*, or as the driver expressing his other–regarding preferences, which is accurately measured by the meter. Once the driver decided how much to give, and where to end the journey, the tester had to ask for a receipt, leave the taxi, and discreetly complete an experimental sheet. The sheet included subjective characteristics of the driver, such as his age, gender (1 if male) and ethnicity, measures of the field including traffic intensity (recorded on a 10 point scale: 1 if Not Busy, 10 if Very Busy) and the weather (1 if raining), and finally characteristics of the ride including whether the driver attempted a conversation (1 if yes), if he offered a cashpoint (1 if yes) and (in the *Reputation* treatment) if he gave a business card or not (1 if one was given). Most importantly, the testers had to record the final meter reading and if the driver completed the journey or not.[14] We present these measures in Table 4.5.

At this stage, it is worth pointing out what the experimental procedure was not. The procedure was not an attempt to obtain free journeys by demanding them from the driver, nor did the testers manoeuvre the driver into making a decision he did not want to take. The testers were instructed to respect the driver at all times, and at no point did the testers question the drivers' right to charge the metered fare. As the tester requests the reduction of the fare, the driver clearly possesses the right to grant or refuse the request and charge the metered amount: the interaction cannot be interpreted as a negotiation.

4.5 Results

In this section, we outline the experimental results. A number of common features are present throughout the analysis. Where non-parametric tests are utilised, both the p–value and test statistic are presented in parentheses. Unless otherwise stated, all tests are two–sided, and in all regressions journeys from all treatments are pooled.

4.5.1 Journey calibration checks

Some initial calibration checks are conducted in order to examine if our expected fare calculations are accurate. Table 4.6 outlines the recorded fare, expected fare and amounts given as a percentage of the expected fare, from journeys where the driver completed the journey. Observations are disaggregated by *Short* and *Long* distance journeys. By comparing the observed fare in a completed journey to its expected fare, the accuracy of our expected fare calculations can be examined. Minor discrepancies

[14] This cannot be inferred from the receipts, which only contain information about the amount paid by the tester.

between recorded and expected fares are to be expected, largely due to variations in traffic intensity and other random shocks.

Formally comparing the recorded and expected fares, no significant differences in the *Short* distance treatment ($p = 0.652$, Sign Test) or *Long* distance treatment ($p = 0.524$, Sign Test) are reported. The amount given as a percentage of the expected fare is not significantly different to the planned 20% in both the *Short* ($p = 1$, Sign Test) and *Long* ($p = 1$, Sign Test) distance treatments. We conclude that our journey planning is accurate.

	Short Distance	*Long Distance*
Recorded Fare (£)	£5.44	£10.42
	(1.29)	(1.465)
Expected Fare (£)	£5.40	£10.02
	(1.07)	(0.781)
Amount Given as a % of the expected fare	27.5%	24.1%
	(0.254)	(0.148)
Completed Journeys	44	22

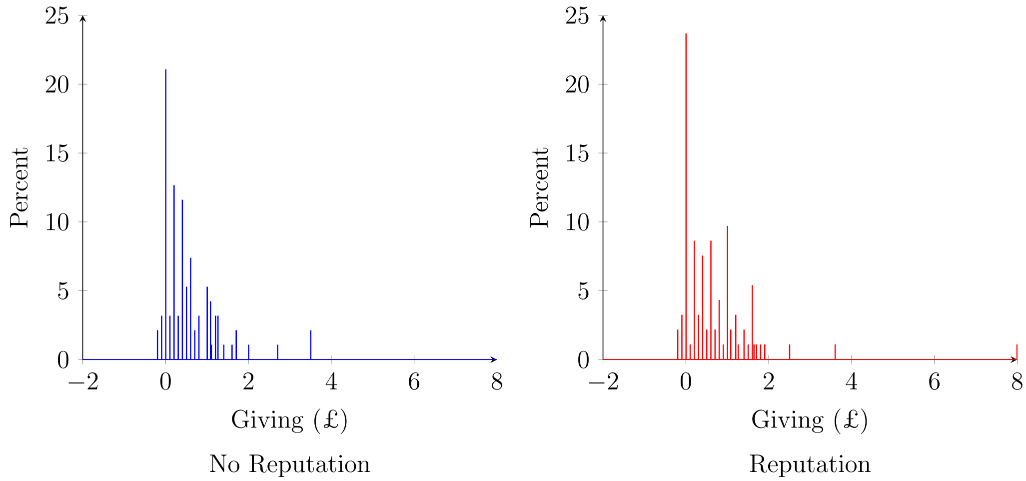
Note: We exclude from these calculations 18 observations where the driver completed the journey, but switched off the meter before the journey was completed. In these 18 cases, we approximate the meter reading by the expected fare. Standard deviations in parentheses.

Table 4.6: Fares, Expected Fares and Average Giving conditional on the Driver Completing the Journey

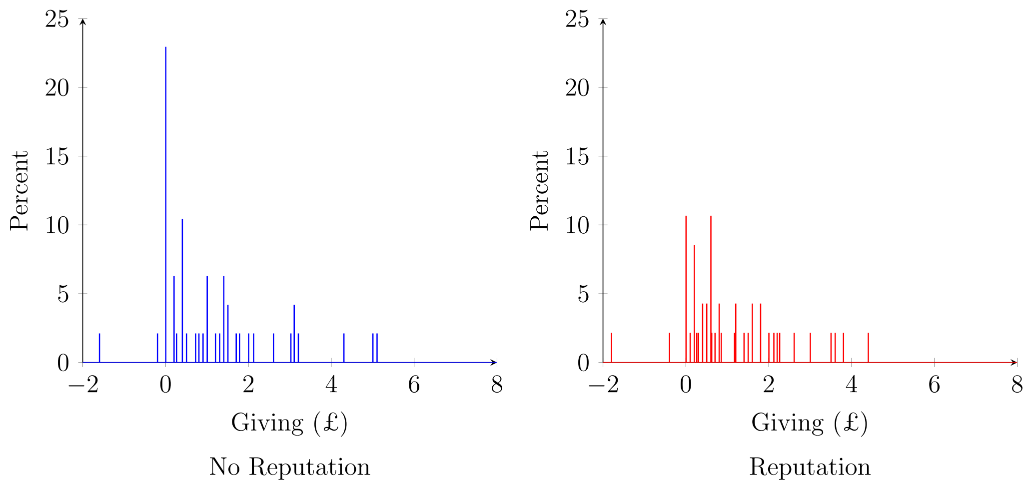
4.5.2 Other regard and reputation effects

Table 4.7 outlines average amounts given by drivers and the proportion of journeys they completed, by treatment. To examine if relative payoffs are a motivating factor behind the amounts that drivers are giving, giving as a percentage of the expected fare is also reported. Figure 4.1 displays the distribution of giving across treatments.

Table 4.8 reports a number of random effects Tobit regressions. In models (1), (2) and (3) giving in pounds by driver i to tester j is the dependent variable. In models (4), (5) and (6), giving as a percentage of the expected fare by driver i to tester j is the dependent variable. Considering giving in this way enables us to control for the variation in journey lengths, and therefore variation in the expected fares of journeys, both within and between treatments. In each regression, dummy variables for the *Long* distance treatment and the *Reputation* treatment are included along with their interaction; the *Short* distance *No Reputation* treatment is taken as the baseline.



(a) Short Distance Journeys



(b) Long Distance Journeys

Figure 4.1: Distribution of Giving, by Treatment

	No Reputation		*Reputation*	
	Short	*Long*	*Short*	*Long*
Amount Given (£)	£0.56	£1.11	£0.71	£1.07
	(0.69)	(1.39)	(1.06)	(1.21)
Amount Given as a % of the Expected Fare	10.6%	11.2%	13.4%	10.5%
	(0.128)	(0.144)	(0.207)	(0.12)
Proportion of Journeys Completed	0.27	0.27	0.31	0.34
Number of Journeys	95	48	93	47

Note: Standard deviations in parentheses.

Table 4.7: Average Driver Giving, by Treatment

	Random Effects Tobit Regressions					
Dep. Variable:	Amount Given (£)			Amount Given as a % of the Exp. Fare		
	(1)	(2)	(3)	(4)	(5)	(6)
Long	0.536**	0.63**	0.609**	-0.005	0.016	-0.002
	(0.217)	(0.248)	(0.245)	(0.062)	(0.071)	(0.07)
Rep.	0.097	0.074	0.091	0.039	0.035	0.04
	(0.154)	(0.156)	(0.154)	(0.044)	(0.045)	(0.045)
Rep. × Long	0.01	0.041	0.025	0.011	0.021	0.017
	(0.272)	(0.276)	(0.271)	(0.078)	(0.079)	(0.078)
Constant	0.68***	0.82***	0.667**	0.178***	0.181***	0.142**
	(0.203)	(0.267)	(0.273)	(0.049)	(0.068)	(0.066)
City Controls	✓	✓	✓	✓	✓	✓
Field Controls		✓	✓		✓	✓
Ride Controls			✓			✓
Observations	283	282	281	283	282	281

Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. The number of observations fall slightly as more controls are included due to missing entries. Models (1), (2) and (3) are left censored at 0, and right censored at the difference between the expected fare had the driver completed the journey, and the amount paid by the tester. Models (4), (5) and (6) are left censored at 0, and right censored at 1.

Table 4.8: Treatment Effects

In each subsequent model, the number of explanatory variables is increased to examine the robustness of the estimated treatment effects. The additional variables we use were those recorded by the testers, outlined in Table 4.5, which we group into 3 distinct sets: Field, City and Ride controls. The set of Field Controls includes the variable for traffic intensity (recorded on a 10 point scale: 1 if Not Busy, 10 if Very Busy), and a dummy controlling for the weather conditions (1 if raining). City Controls includes dummies for the journey taken in Birmingham, Trafford or Salford (1 if yes), with those taken in Manchester taken as the baseline. Ride Controls includes dummies controlling for whether the driver offered to take the passenger to a cash–point (1 if offered) and if he tried to engage in a conversation (1 if yes).

Result 1. *The majority of taxi drivers give at least part of the journey for free.*

Support. Considering journeys from the *No Reputation* treatment, the null hypothesis of no giving can be rejected at the 1% level in both *Long* and *Short* distance journeys

($p < 0.01$, both cases, Sign Test). Over 70% of drivers give part of the journey for free, and over 25% of all journeys were completed in full. Parametric support is given in Table 4.8, with a positive and significant constant in all regression models ($p < 0.05$, in all cases). Similar findings are observed in the *Reputation* treatment, with over 75% of drivers giving at least part of the journey for free, and 32% of all journeys being completed in full.

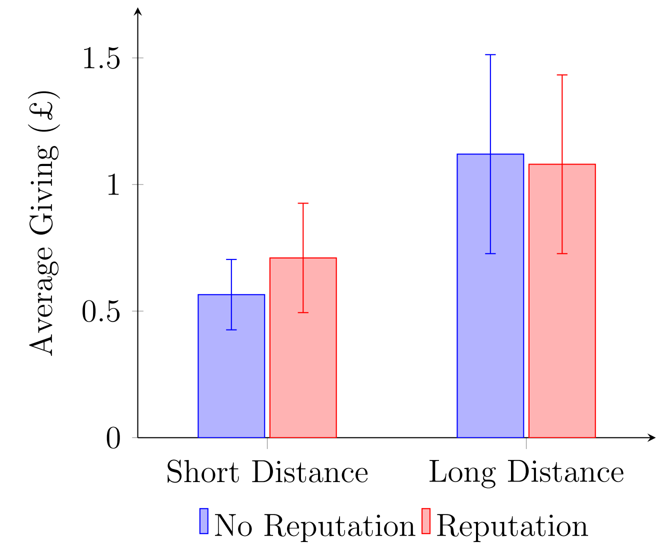
Result 2. *Driver giving is proportional to the distance of the journey.*

Support. Examining journeys from the *No Reputation* treatment, average driver giving is significantly different in *Short* distance journeys in comparison to *Long* distance journeys ($p = 0.056$, Robust Rank Order Test). This is shown graphically in Figure 4.2a. The distribution of giving is also found to vary by the distance of the journey ($p = 0.039$, Kruskal–Wallis Test). Table 4.8, regressions (1), (2) and (3) support these conclusions, reporting significant and positive coefficient estimates on the *Long* distance dummy ($p < 0.05$), whilst the coefficient on the *Reputation* dummy alone is not significant ($p > 0.1$). However, when giving as a percentage of the expected fare is considered, no significant differences are reported by distance ($p = 0.86$, Robust Rank Order Test) (see Figure 4.2b). Further, the distance of the journey has no significant effect on its distribution ($p = 0.86$, Kruskal–Wallis Test). Estimates from Table 4.8 models (4), (5) and (6), support this conclusion; no significant treatment effects are reported when the dependent variable is giving as a percentage of the expected fare ($p > 0.1$ in all cases, in all regressions), suggesting giving is proportional to the length of the journey, and therefore the amount the driver can give.

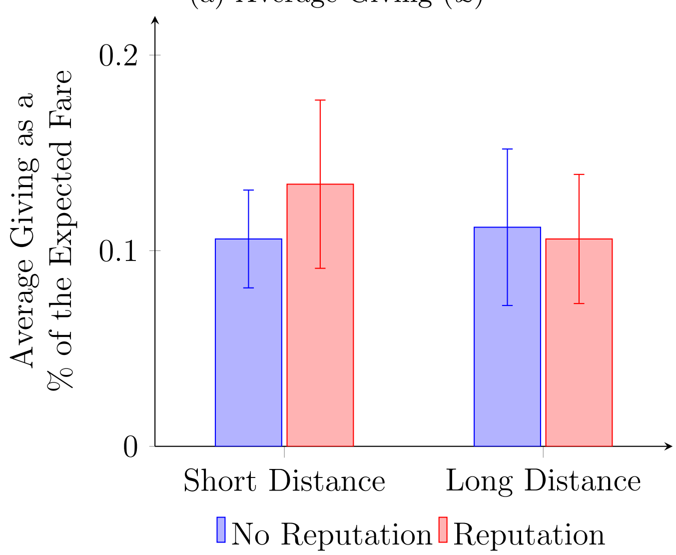
Results 1 and 2 suggest that taxi drivers have other–regarding preferences that appear to be well defined over the relative payoff between themselves and the passenger. These results support the idea that such other–regarding behaviour can, and does, exist within competitive market settings. The effect of other–regarding preferences on the market is clear: the drivers' other–regarding preferences lower the price of taxi journeys.

Result 3. *Reputational concerns do not explain the extent of giving.*

Support. Comparing average giving between *Reputation* and *No Reputation* treatments, no significant differences are reported in either *Short* or *Long* distance journeys



(a) Average Giving (£)



(b) Average Giving as a Percentage of the Expected Fare

Note: Vertical bars represent 95% confidence intervals.

Figure 4.2: Average Giving

($p = 0.34$ and $p = 0.67$, Robust Rank Order Tests). Similarly, reputational concerns have no significant impact on the distribution of giving in either *Long* or *Short* distance treatments ($p = 0.44$ and $p = 0.67$, Kruskal–Wallis Test). The same is true for giving as a percentage of the expected fare, with no significant differences found between *Reputation* and *No Reputation* treatments in *Short* or *Long* distance journeys, or when journeys are pooled ($p > 0.1$ in all cases, Robust Rank Order Tests). Estimates from Table 4.8 supports these results, with the coefficient on the *Reputation* dummy found to be not significant at conventional levels across regressions ($p > 0.1$ in all cases).

Result 3 outlines how the drivers' behaviour is, on average, unaffected by reputational concerns. However, it is possible that the effect of the *Reputation* treatment on the drivers' behaviour could either promote or diminish other–regard. It may promote behaviour if the drivers believe their other–regard will increase the probability of being contacted for future journeys by the passenger, or that their other–regard might be reciprocated in future journeys through tipping. Alternatively, it may diminish giving if drivers do not want a repeated interaction with a passenger who asks for a portion of the fare for free, especially if they suspect the passenger of using this trick in order to induce drivers to behave in an other–regarding manner.

In addition, the drivers' behaviour may depend on the appearance characteristics of the tester. To explore this further we examine driver giving conditional on the testers' ethnicity. Summary statistics are given in Table 4.9 and Figure 4.3 displays the proportion of completed journeys by tester ethnicity graphically. To determine the effect of the testers' ethnicity on driver giving, Table 4.10 outlines the results from a number of random effects Tobit regressions. In each case, giving in pounds by driver i to tester j is the dependent variable. The estimated coefficients on a dummy controlling for whether the tester was black (1 if yes), South–Asian (1 if yes) and if they were male (1 if yes) are reported; white testers are taken as the baseline.

To examine the robustness of the estimated coefficients, in each model we systematically increase the number of explanatory variables, which are grouped into 6 sets: Treatment, Driver, Tester, Ride, Field and City Controls. Treatment controls include dummies for each of the treatments (1 if *Long*, and 1 if *Reputation*) and the interaction, Driver controls include the driver's age and gender (1 if male). Tester controls include the tester's gender (1 if male), which is reported, and also their age. Field, Ride and City controls are identical to those described for Table 4.8. For each tester, we also include their average rating for each appearance characteristic: aggressiveness, attractiveness, friendliness, trustworthiness and wealthiness. The estimated coefficients on these variables are included in Table 4.10.

Result 4: *Drivers give the least to black testers.*

Support. Pairwise comparisons of average giving by drivers to white, black and

		Treatment				
		No Reputation		*Reputation*		
Ethnicity		*Short*	*Long*	*Short*	*Long*	Total
White	Amount Given (£)	£0.60	£1.21	£0.85	£1.2	
		(0.564)	(1.401)	(0.725)	(1.29)	
	Amount Given, % of Exp. Fare	11.1%	12.6%	16.3%	12%	
		(0.107)	(0.157)	(0.139)	(0.129)	
	Journeys	60	26	49	29	164
Black	Amount Given (£)	£0.26	£0.79	£0.57	£1.05	
		(0.396)	(0.991)	(1.57)	(1.312)	
	Amount Given, % of Exp. Fare	5.1%	7.6%	11.3%	9.9%	
		(0.08)	(0.09)	(0.309)	(0.125)	
	Journeys	26	11	30	11	78
South–Asian	Amount Given (£)	£1.23	£1.22	£0.52	£0.54	
		(1.402)	(1.76)	(0.654)	(0.526)	
	Amount Given, % of Exp. Fare	22.9%	11.3%	8.2%	5.4%	
		(0.255)	(0.16)	(0.115)	(0.052)	
	Journeys	9	11	14	7	41
Total		95	48	93	47	283

Note: Standard deviations given in parentheses.

Table 4.9: Summary Statistics, by Tester Ethnicity.

South–Asian testers in the *No–Reputation* treatment reveals no significant differences between white and South–Asian testers in the *Short* or *Long* distance treatments $p = 0.64$ and $p = 0.46$, Robust Rank Order Tests). However, significant differences between white and black testers are reported in the *Short* but not in the *Long* distance treatment ($p = 0.001$ and $p = 0.39$, Robust Rank Order Tests). Similarly, a significant difference between South–Asian and black testers is found in the *Short* but not in the *Long* distance treatment ($p = 0.06$ and $p = 0.47$, Robust Rank Order Tests). Considering giving by the amount given as a percentage of the expected fare reveals that both white and South–Asian testers are given significantly more than black testers ($p = 0.005$, $p = 0.025$, Robust Rank Order Tests), but no differences are found between white and South–Asian testers ($p = 0.31$, Robust Rank Order Test). The estimates in Table 4.10 further support the non–parametric results: across all regressions, the coefficient on the black dummy is negative, highly significant ($p < 0.01$, Wald Tests), and robust to changes in the model specification.

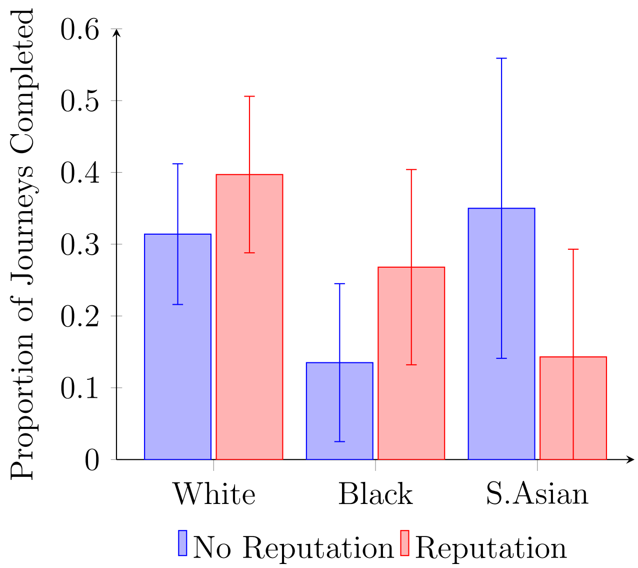
The differential treatment of testers by ethnicity remains in the *Reputation* treatment, with white testers receiving more than black testers in the *Short* distance treat-

Random Effects Tobit Regressions					
Dep. Variable:	Amount Given (£)				
	(1)	(2)	(3)	(4)	(5)
Black	-0.645***	-0.634***	-0.612***	-0.585***	-0.695***
	(0.197)	(0.191)	(0.306)	(0.187)	(0.179)
South–Asian	-0.132	-0.261	-0.202	-0.252	-0.006
	(0.264)	(0.241)	(0.202)	(0.236)	(0.238)
Male		-0.334*	-0.324*	-0.32*	-0.419*
		(0.172)	(0.149)	(0.172)	(0.228)
Aggressiveness					0.104
					(0.214)
Attractiveness					0.179
					(0.109)
Friendliness					-0.094
					(0.109)
Trustworthiness					0.087
					(0.252)
Wealthiness					-0.175
					(0.126)
Constant	0.491	1.18	1.02	1.34*	1.184
	(0.678)	(0.752)	(0.715)	(0.765)	(2.65)
Treatment Controls	✓	✓	✓	✓	✓
Driver Controls	✓	✓	✓	✓	✓
Tester Controls		✓	✓	✓	✓
Ride Controls			✓	✓	✓
Field Controls				✓	✓
City Controls				✓	✓
Observations	275	275	274	274	274

Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. The number of observations falls slightly as more controls are included due to missing entries. All models are left censored at 0, and right censored at the difference between expected fare, had the driver completed the journey, and the amount paid by the tester.

Table 4.10: The Determinants of Driver Giving

ment ($p < 0.001$, Robust Rank Order Test) although no difference is observed between white and South–Asian testers ($p = 0.63$, Robust Rank Order Tests). No differences are reported between black and South–Asian testers in either distance treatment ($p > 0.1$ in both cases). Comparing giving as a percentage of the expected fare reveals differences in giving between white and black and white and South–Asian testers ($p < 0.001$ and $p = 0.003$, Robust Rank Order Tests), but no difference between black and South–Asian testers ($p = 0.9$, Robust Rank Order Test).



Note: Vertical bars represent 95% confidence intervals.

Figure 4.3: Proportion of Journeys Completed, by Tester Ethnicity

The proportion of completed journeys, by tester ethnicity, is now considered. Table 4.11 reports a number of random effects Probit regressions, where the dependent variable is a dummy that takes a value of 1 if the journey was completed. We increase the number of explanatory variables in each subsequent model, and use the same control variables as outlined in Table 4.10.

Result 5: *Drivers are least likely to complete a journey for a black tester.*

Support. Comparing the proportion of journeys that were completed, by tester ethnicity, black testers have their journey completed significantly less often than white and South–Asian testers in the *No Reputation* treatment ($p = 0.045$ and $p = 0.088$, Fisher's Exact Test). No significant differences are reported between white and South–Asian testers ($p = 0.793$, Fisher's Exact Test). The results from the random effects Probit regressions in Table 4.11 outline how the estimated coefficient on the black dummy is negative and significant ($p = 0.05$). This estimate is robust to specification changes, and becomes increasingly significant as more controls are included. Similar to the coefficient estimates in Table 4.10, none of the appearance characteristics are significant, except the appearance of wealthiness ($p = 0.06$), which has a negative effect: more

Random Effects Probit Regressions					
Dep. Variable:	Journey Completed				
	(1)	(2)	(3)	(4)	(5)
Black	-0.500**	-0.506**	-0.531**	-0.517**	-0.643**
	(0.196)	(0.21)	(0.213)	(0.221)	(0.242)
South–Asian	-0.309	-0.348	-0.348	-0.414	-0.053
	(0.240)	(0.244)	(0.245)	(0.268)	(0.321)
Male		-0.281	-0.271	-0.258	-0.318
		(0.182)	(0.185)	(0.196)	(0.309)
Aggressiveness					-0.195
					(0.336)
Attractiveness					0.207
					(0.146)
Friendliness					-0.198
					(0.24)
Trustworthiness					-0.036
					(0.33)
Wealthiness					-0.321*
					(0.171)
Constant	-0.906	-0.06	-0.140	0.288	3.41
	(0.760)	(0.847)	(0.861)	(0.909)	(3.51)
Treatment Controls	✓	✓	✓	✓	✓
Driver Controls	✓	✓	✓	✓	✓
Tester Controls		✓	✓	✓	✓
Ride Controls			✓	✓	✓
Field Controls				✓	✓
City Controls				✓	✓
Observations	275	275	274	274	274

Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level.

Table 4.11: Determinants of Journey Completion

wealthy looking testers are less likely to have a journey completed.

Results 4 and 5 outline how black testers are treated significantly worse than white and South–Asian testers. As the coefficient on trustworthiness is insignificant, and its direction the opposite we would expect, statistical discrimination is likely not the explanation. Status is also unlikely to be a factor, as wealthiness has a negative effect on giving and our black testers are rated as appearing the *least* wealthy as outlined in Section 4.4.1. Indeed, the inclusion of the appearance characteristics increases the magnitude of the coefficient of the black dummy in both the Tobit and Probit regressions. The evidence points towards taste based discrimination.

Result 6: *Reputational concerns increase driver giving when the tester is white, have no effect when the tester is black and reduce giving when the tester is South–Asian.*

Support. White testers are given significantly greater amounts as a percentage of the expected fare in the *Reputation* treatment compared to the *No Reputation* treatment ($p = 0.06$, Robust Rank Order Test). They also receive significantly more in absolute terms as a result of reputation in the *Short* distance treatment ($p = 0.053$, Robust Rank Order Test), although no significant difference is observed in the *Long* distance treatment ($p = 0.61$, Robust Rank Order Test). Black testers see no significant differences as a result of reputation ($p > 0.1$ in all cases, Robust Rank Order Tests). South–Asian testers see no effect of reputation on absolute giving in both the Short ($p = 0.27$, Robust Rank Order Test) and *Long* distance treatments ($p = 0.5$, Robust Rank Order Test), and a (weakly) negative effect is reported in giving as a percentage of the expected fare ($p = 0.15$, Robust Rank Order Test) resulting from driver reputational concerns.

Result 6 can be explained by drivers' beliefs about their expected payoffs from their future interaction with the passenger, and is unlikely to be due to beliefs that white passengers are most able to contact them: drivers give business cards uniformly across all tester ethnicities ($p > 0.1$ in all comparisons, Fisher's Exact Tests). There are, however, two different belief channels through which the disparity can occur, either through drivers' beliefs about the probability of a repeated interaction, or through their beliefs about their earnings from a repeated interaction. Drivers may believe the probability of a future interaction is greatest for a white passenger, or that by expressing other–regard they increase this probability by more than if the tester was black or South–Asian. Alternatively, drivers may believe white passengers are more likely to reciprocate their other–regard in a future interaction through tipping, as shown by Ayres *et al.* (2005), who report that white passengers in the United States tip approximately twice as much as passengers of other ethnicities.

4.5.3 Structural models

The reduced form estimates provided in Section 4.5.2 provide evidence of variation in driver giving that is conditional on the testers ethnicity. However, they do not provide quantitative estimates of the preferences underlying this behaviour. We now estimate the parameters of a number of utility functions, in order to link our empirical analysis to behavioural theory. We work under the assumption that drivers are 'narrowly bracketing' their decisions, and thus, are not taking into account the annual income of the passenger or themselves (Read *et al.*, 1999).

To begin, it is assumed that each driver has distributional preferences over their own payoff, m, and the passenger's payoff, y. For a given journey, the driver's payoff is equal to the amount paid by the passenger, $s \in \{4, 8\}$, minus the amount of journey he gives them for free, $x \in [0, \bar{x}]$, and minus the fuel costs associated with the entire journey, $g(x) \cdot p$, where $g(x)$ is the distance of the journey in miles, and p the price in fuel per mile travelled: $m = s - x - g(x) \cdot p$. When the driver selects $x = 0$, he stops when the meter reaches the amount the passenger can afford; $x = \bar{x}$ implies he completed the journey. The passenger's payoff is defined as being equal to the amount given to her by the driver, x, so $y = x$. As the appearance characteristics of the tester are not found to be significant determinants of amounts given at the 5% level, as outlined in Table 4.10, we exclude these from the structural model.[15]

The distance driven by the driver for each journey is approximated using the final meter reading and corresponding fare table for each local authority, and we assume there were no wait times. For each journey we calculate the drivers' fuel costs conditional on the traffic intensity, as reported by the tester, and use fuel costs per mile based on the fuel efficiency of the LTI TXII Hackney Carriage.[16]

We incorporate traffic intensity into the model as traffic flows will affect a driver's fuel costs, with a higher traffic intensity forcing the driver to break more often, or drive in a lower, less fuel–efficient gear. When traffic intensity is reported below the median of 4, we assume fuel efficiency to take a high *extra–urban* rate of 42 miles per gallon (£0.12 per mile), an *urban* rate of 29 miles per gallon when it is below average (£0.17 per mile) and a *combined* rate of 36 miles per gallon when it is equal to the average (£0.14 per mile).[17,18] The price of fuel is taken to be £1.10 per litre, the average price of diesel at the time the experiment took place, which is assumed to be identical across drivers.

[15] Pearson's correlation coefficients reveal the following correlations with the amount given and appearance characteristics: aggressiveness, $r = -0.03$, attractiveness, $r = 0.09$, friendliness, $r = 0.02$, trustworthiness, $r = 0.05$ and wealthiness $p = 0.06$. None are significant at conventional levels ($p > 0.1$ in all cases).

[16] This model of taxi is chosen as it is the most common amongst the drivers we surveyed, as shown in Table 4.14 in Section 4.7. In reality, there are only small differences in fuel efficiency between models.

[17] Our estimates are quantitatively robust to changes in how traffic affects the drivers' fuel costs.

[18] Fuel efficiency figures are taken from http://www.fuel-economy.co.uk/mpg.php.

Model	Functional Form	Description	Reference
(1)	$u(y,m) = my^{\theta}$	Cobb-Douglas	Cox *et al.* (2007)
(2)	$u(y,m) = m + \theta y^{\alpha}$	Inequity Aversion ⋄	Fehr & Schmidt (1999)
(3)	$u(y,m) = (m^{\alpha} + \theta y^{\alpha})\alpha^{-1}$	CES◇	Cox *et al.* (2007)

⋄ When $\alpha = 1$, both models (2) and (3) are identical to the Fehr & Schmidt (1999) model of inequality aversion.
◇ Constant Elasticity of Substitution.

Table 4.12: Estimated Functional Forms

Table 4.12 outlines the three functional forms of utility that we estimate. Due to the nature of the driver's choice, the forms estimated are limited to one and two parameter specifications. Across specifications, parameter θ represents the other-regarding preference parameter, or the utility weight that the driver places on the payoff of the passenger. Parameter α, in specifications (2) and (3), is a convexity parameter. In all cases, when $\alpha = 1$, utility is linear. The specification of Cox *et al.* (2007) in models (1) and (3) are chosen because in these functions drivers' preferences are homothetic: preferences over relative payoffs are well defined, and our data suggests drivers have such preferences. Model (3) is particularly flexible, as outlined by Cox *et al.* (2007). A generalised form of the Fehr & Schmidt (1999) inequity averse function is selected in model (2) due to its prominence in the literature: incorporating a convexity parameter will allow us to examine if utility is linear in own and others' payoffs, as is often assumed.

In each specification, following Chen & Li (2009), ethnic identity is incorporated into the model by assuming that other–regarding preferences, θ, are group contingent, and that these preferences are a function of the ethnic identities of the driver and tester. We specify θ as the following function,

$$\theta = \bar{\theta} \cdot (1 + a \cdot m_1 + b \cdot m_2 + c \cdot m_3 + d \cdot m_4 + e \cdot m_5) + \epsilon, \tag{4.1}$$

where m_i are dummy variables that take values of 1, conditional on the driver's and passenger's ethnicity; m_1 and m_2 take values of 1 when the driver is white, and when the passenger is black or South–Asian respectively; m_3, m_4 and m_5 take values of 1 when the driver is South–Asian, and when the passenger is white, black or South–Asian. We limit the analysis to journeys with white and South–Asian drivers due to the small number of journeys taken with black drivers. Journeys with both a white driver and a white passenger are taken as the baseline. The identity parameters, a, b, c, d and

			Model Specification					
Ethnicities			*Without Identity*			*With Identity*		
Driver	Passenger		(1)	(2)	(3)	(1)	(2)	(3)
		σ	0.652*	0.857**	0.279***	0.634*	0.741*	0.258***
			(0.362)	(0.438)	(0.101)	(0.357)	(0.385)	(0.089)
		$\bar{\theta}$	0.021	0.811***	0.576***	0.287***	1.245***	0.721***
			(0.061)	(0.104)	(0.199)	(0.109)	(0.214)	(0.137)
		α		0.655***	0.84***		0.676***	0.846***
				(0.134)	(0.098)		(0.128)	(0.089)
White	Black	a				-0.781***	-0.244	-0.132
						(0.235)	(0.173)	(0.104)
White	S. Asian	b				0.53	0.114	0.043
						(0.530)	(0.119)	(0.062)
S. Asian	White	c				-0.935***	-0.359*	-0.181
						(0.336)	(0.196)	(0.138)
S. Asian	Black	d				-2.512***	-0.856***	-0.481*
						(0.818)	(0.309)	(0.264)
S. Asian	S. Asian	e				-1.05	-0.37	-0.209
						(0.980)	(0.387)	(0.27)
Observations			132	132	132	132	132	132

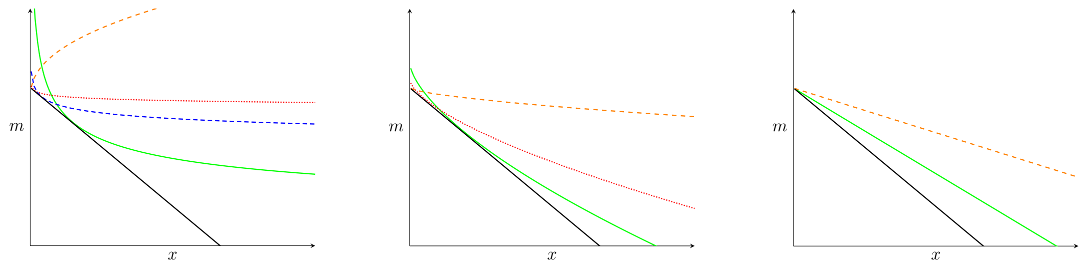
Note: Standard errors clustered at the tester level. Robust standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. Only journeys from the *No Reputation* treatment are used, from both the *Short* and *Long* distance treatments. Journeys where the driver stopped before the meter reached the amount the tester could afford are coded as the driver giving £0. Reduced form estimates that support these results are given in Table C.2 in Appendix C.2.

Table 4.13: Structural Parameter Estimates

e, therefore capture the additional effects of variations in the drivers' and passengers' ethnicity on θ. The function θ is assumed to be identical across drivers, except for an idiosyncratic error term, $\epsilon \sim \mathcal{G}(0, \sigma^2)$, where $\mathcal{G}$ is the type I extreme value distribution. The estimation strategy is outlined in Appendix C.2.

First, we estimate the parameters $\bar{\theta}$, α, and σ, with the following restriction: $a = b = c = d = e = 0$. The results are displayed in Table 4.13 under the *Without Identity* heading. Second, we remove the identity parameter restrictions, and let the model pick their values; the results are displayed in Table 4.13 under the *With Identity* heading. The parameters are estimated using only the journeys from the *No Reputation* treatment to avoid any potential confounding effects originating from the drivers' reputational concerns, with observations clustered at the tester level.[19]

[19]The results are quantitatively similar if the parameters are estimated pooling the observations from both the *No Reputation* and *Reputation* treatment.



— Baseline: $\theta = 0.287$
- - - S.Asian–Black $\theta = -0.434$
······ S.Asian–White $\theta = 0.019$
- - - White–Black: $\theta = 0.063$

(a) Model 1

— Baseline: $\theta = 1.245$
- - - S.Asian–Black $\theta = 0.179$
······ S.Asian–White $\theta = 0.797$

(b) Model 2

— Baseline: $\theta = 0.721$
- - - S.Asian–Black $\theta = 0.374$

(c) Model 3

Note: The passenger's payoff, x, is plotted on the x axis and the driver's payoff, m, is plotted on the y axis. In each panel, the thick black curve represents a hypothetical 'budget line'. In Figure 4.4b, $\alpha = 0.676$, and in Figure 4.4c, $\alpha = 0.846$, for each of the indifference curves, as estimated in Table 4.13. Where identity parameters are found not to be significant, other–regarding preferences are taken to be equal to the baseline.

Figure 4.4: Estimated Indifference Curves

Models (1), (2) and (3) in Table 4.13 each outline how the drivers have other–regarding preferences. In the single parameter specification of model (1), although $\bar{\theta}$ is not significantly different to 0 ($p > 0.1$), the dispersion of preferences, σ, is found to be significant, suggesting many of the drivers do have other–regarding preferences. In models (2) and (3), $\bar{\theta}$ is estimated to be positive and significant at the 1% level, with significant preference heterogeneity reported, with $\sigma > 0$ ($p \leq 0.1$ across models).[20]

Interestingly, when identity is included, the estimates of α, $\bar{\theta}$ and σ remain robust. In those models that include identity, a number of patterns relating to ethnic identity emerge. First, parameter d is reported to be negative and significant, with $p \leq 0.01$ in model (1) and (2), and $p = 0.06$ in models (3). This suggests that South–Asian drivers' other–regarding preferences are significantly smaller when faced with a black passenger, in comparison to both white and South–Asian passengers; in model (1), giving to a black passenger is estimated to be a *bad* for a South–Asian driver (see Figure 4.4a).

Second, weak evidence that white driver preferences are reduced when faced with a black passengers is reported, with $a = -0.781$ ($p < 0.01$) in model (1), although its significance is not robust to specification changes. The parameter measuring the effect of

[20]For comparison, Cox *et al.* (2007) estimate model (3) using dictator game data, with slightly different assumptions regarding ϵ, and report $\theta = 0.417$ and $\alpha = 0.255$.

South–Asian / white interactions, c, is similar, estimated to be negative and significant in models (1) and (weakly) significant in model (2) ($p = 0.06$), but insignificant in model (3) ($p > 0.1$). Finally, no evidence is found that white drivers' preferences are influenced by South–Asian passengers, with b found to be positive, but insignificant in all models ($p > 0.1$ in all cases). Figure 4.4 plots the estimated indifference curves from model (1), (2) and (3) graphically.

4.6 Robustness checks

As we examine the data for heterogeneous treatment effects for different ethnic subgroups, the statistical significance of some of these effects may be an artefact of multiple hypothesis testing. To account for this, we adjust the calculated p-values used to support Results 4, 5 and 6 using the Holm–Bonferroni procedure (Holm, 1979). This procedure is used over the more conservative Bonferroni procedure because of its increased power (Holm, 1979; List *et al.*, 2016). We first consider the robustness of the p–values calculated from non–parametric testing, and then those obtained from the regression analysis.

For each result in Section 4.5, Table C.3, given in Appendix C.2, presents the unadjusted and Holm–Bonferroni adjusted p–values for each hypothesis tested given the 'family' of hypotheses each test falls into. Similar to List *et al.* (2016), we define the 'family' of hypotheses as the group of tests related to a particular outcome compared within a treatment, or the group of tests related to a particular outcome compared between treatments.

Table C.4, also given in Appendix C.2, presents the adjusted p–values for the hypothesis tests conducted on the *Black*, *South–Asian* and *Male* dummies from each of the regression models in Table 4.10 and Table 4.11 in Section 4.5. To adjust the p–values, the family of hypotheses is defined as the number of variables of interest tested for significance within each regression, given in the final column as m. We include within the family of tests, where appropriate, ethnicity, gender and appearance characteristics.

The adjusted p–values in Table C.3 and C.4 provide a number of insights. First, the negative differential between giving to black and white testers concluded in Result 4 is robust: both the non–parametric and parametric results are robust to adjustments for multiplicity (Hypotheses 1, 7, 10, 16, 17, Table C.3 and Hypotheses 1–5, Table C.4). The difference in giving between South–Asian and black testers is reasonably robust, but only remains significant when all observations from both the *Short* and *Long* distance treatments are pooled (Hypothesis 9, Table C.3). Second, although the non–parametric results in support of Result 5 are not found to be significant once adjusted (Hypotheses 10–12, Table C.3), the parametric results are found to be robust (Hypotheses 19–21, Table C.4). However, Result 6 does not appear to be as robust as Results 4 and 5 (Hypotheses 22 and 28, Table C.3).

4.7 Discussion

Three main questions arise from the results in Section 4.5: (1) can the extent of giving be explained by the drivers finding a convenient location to stop?; (2) can earnings expectations stemming from bargaining with passengers explain the drivers' behaviour?; (3) can social pressure explain the extent of giving?

To examine questions (1) and (2) we conducted a complementary survey of 50 taxi drivers from ranks used within the study, 65 passengers that were queuing for a taxi, and observed the behaviour of 97 passengers entering taxis from a rank.[21] To address (1) we asked drivers the number of daily journeys they take, how many of these journeys are from taxi ranks and what they believe the average fare is. Drivers were also asked about the expected fare of an example *Short* and *Long* distance journey, where the example journeys were journeys that we used within the study. They were asked if they would be willing to bargain over the journey specified *before* the journey began, and the lowest fare they would accept if they were willing. In addition, they were asked if they would be willing to bargain with a passenger who was inside the taxi.[22] Finally, we asked them what they did upon completing a journey using a multiple choice question: return to a home rank, return to a different rank, cruise and look for a passenger, or do something else. Passengers were asked if they ever bargained with the driver when catching a taxi from the rank.

The drivers' responses are presented in Table 4.14, Panel A, and the passenger responses and the observation results are presented in Panel B.

The responses in Table 4.14 highlight two main points relating to (1). First, the vast majority of taxi journeys are taken from ranks (92%). This suggests that giving to the passenger, by continuing to drive away from the rank, is not done at the drivers'

[21]The survey and observations were conducted in Manchester. The questionnaire is given in Appendix C.1.

[22]Drivers were also asked to report their income, but the majority refused to disclose this information.

Panel A:	Driver Survey, $N = 50$	
	No. daily journeys	12
		(4.4)
	No. journeys that start at a rank	11
		(4.47)
	Average fare (£)	6.41
		(1.4)
	Modal Taxi Model	LTI TXII
Short distance journeys	Expected fare (£)	6.17
		(0.778)
	Willing to bargain? (1 if yes)	0.06
		(0.242)
	Lowest fare if willing (£)	4.73
		(2.11)
	Willing to bargain inside the taxi? (1 if yes)	0.04
		(0.2)
	Upon completion	Return to home rank (73%) Return to a diff. rank$^\diamondsuit$ (16%) Cruise (10%)
Long distance journeys	Expected fare (£)	11.85
		(1.97)
	Willing to bargain? (1 if yes)	0.12
		(0.328)
	Lowest fare if willing (£)	9.33
		(0.328)
	Willing to bargain inside the taxi? (1 if yes)	0.04
		(0.2)
	Upon completion	Return to home rank (76%) Return to a diff. rank$^\diamondsuit$ (10%) Cruise (10%)
Panel B:	Passenger Survey	
	Do you bargain? (1 if yes), $N = 65$	0.03
		(0.181)
	Observed bargaining (1 if yes), $N = 97$	0.01
		(0.1)

Note: All responses relate to journeys taken between 9am–5pm. Standard deviations in parentheses.
$^\diamondsuit$ The majority of drivers specifying this response outlined that they would return to different rank in the centre of the city.

Table 4.14: Driver and Passenger Survey Responses

convenience. On the contrary, driving away from the rank is the same as driving away from the next passenger, and therefore is costly. Second, only 10% 'Cruise' upon completing a journey, with the vast majority returning to a home rank and only a minority returning to a different rank: drivers reported returning to busy city ranks.

This is clearly because that is where the passengers are. Further, as drivers are given ample time to stop, 975 yards ($\sim$ 1 kilometre) in the *Short*, and 1950 yards ($\sim$ 2 kilometres) in the *Long* distance treatment, it seems unlikely they continue out of convenience. There is even less reason to think the distance required to find a convenient location to stop is proportional to the length of the journey.

In relation to (2), from Table 4.14 note that only 6% of drivers said they would bargain with a passenger before the passenger was inside the vehicle for the *Short* distance journey, and only 12% in the *Long* distance journey; the lowest fare they would accept is also above the amount our testers could afford. The majority would refuse to bargain with them prior to the journey beginning, and only 2 reported they would bargain with a passenger mid–journey. Their expected fare estimates are also in–line with our own calculations. Our survey and observation of passengers also shows the desire to negotiate is limited, with only a single passenger observed attempting to bargain with a driver and only 2 reporting that they did bargain with drivers over fares. Therefore, it seems unlikely that driver giving is the result of earnings expectations stemming from passenger bargaining, as the vast majority of journeys are not bargained over.

Question (3) implies that drivers are concerned about appearing unkind to the passenger, and give despite having a preference not to. This would resonate with the conclusion of DellaVigna *et al.* (2012). However, *not* giving away goods and services for free in a market setting is unlikely to be perceived as unkind. This contrasts with charitable giving, where giving to those who need it might be viewed as a normative action. Further, in the context of our study, passengers could easily have taken an alternative and cheaper mode of transport, or could have walked the final portion of the journey they couldn't afford.[23]

4.8 Conclusion

We report evidence that the majority of taxi drivers express other–regarding preferences in transactions completed in a competitive market setting, and find little evidence of the reputational concerns that are often used to explain such behaviour. Our conclusions contrast with the results of previous prominent field experiments and standard

[23] A passenger would have to walk $\sim$ 1 kilometre in the *Short* distance treatment, and $\sim$ 1.8 kilometres in the *Long* distance treatment to complete the journey if they exited the taxi at the amount they could afford.

economic theory, but resonate with the results of numerous laboratory experiments and behavioural theories of social preferences. Within a highly competitive market setting, we observe individuals behaving altruistically.

Variation in the ethnicity of the driver and the tester also allows us to explore recent theories of discrimination, namely, that other–regarding preferences are group–contingent. We find strong evidence that the drivers' propensity to give is significantly smaller when the passenger is black. This result is robust to controlling for variation in the testers' appearance, variation that may otherwise be driving the result. Parameter estimates from a number of structural models reveal that white and South–Asian drivers' other–regarding preferences are group–contingent, being significantly smaller when faced with a black passenger. Weaker evidence that South–Asian drivers' preferences are reduced when faced with a white passenger are also reported.

The effect of reputation on drivers' behaviour is also found to be conditional on the ethnic identity of the passenger. When the passenger is white, drivers behave significantly more other–regarding, and give significantly more of the ride for free. No such result is found for black and South–Asian passengers, with a weakly negative effect of reputation on giving to South–Asians. The potential of a repeated interaction also fails to remove the differential treatment of testers conditional on their ethnic identity. This suggests that drivers' beliefs about the behaviour of individuals also varies with identity.

We acknowledge that markets where transactions are automated or done through a computer, such as asset and financial markets, are unlikely to see the types of behaviour observed here. This is because the nature of the interaction between buyer and seller does not allow for such preferences to be expressed, as market agents are not given the opportunity to behave in such a manner. However, many other types of markets exist. In markets where bilateral face to face interactions are common place we might expect other–regarding preferences to play a much greater role than previously suggested.

Chapter 5

Conclusion

The purpose of this thesis has been to examine the behavioural theory that social preferences are group–contingent, or that the differential treatment of others could stem from discrimination being manifested within individuals' preferences for the welfare of others. The chapters in this thesis employ laboratory and field experiments in order to consider a range of other–regarding preferences, in a multitude of strategic settings and in both market and non–market frameworks. In each chapter we complement our findings by structurally modelling subjects' underlying preferences, directly linking our empirical findings to the theories we wish to test. Although we find no evidence of such discriminatory tastes in Chapter 2, Chapters 3 and 4 present compelling evidence that is consistent with the group–contingent social preference hypothesis.

A number of conclusions can be drawn from our results, each having implications for future avenues of research. The first is that individuals' identities are multifaceted and can influence behaviour in complex ways. This likely explains why no discriminatory behaviour is observed in Chapter 2 and almost certainly explains the various ethnic interaction effects observed in Chapter 4. It could also explain the strong negative out–group discrimination observed in Chapter 3, which contrasts to the positive in–group discrimination typically observed, possibly a consequence of the social context in which the experiment is conducted. Future research should be mindful of the ways in which natural identities interact, particularly when studying diverse subject pools, and carefully consider how these interactions might be influenced by the context in which they are being studied.

A second conclusion drawn from this thesis is the importance of institutions in determining when other–regarding behaviours and discrimination are likely to play significant roles. For example, as highlighted by Heckman (1998), economists often (incorrectly) assume that taste–based discrimination is irrelevant in market settings, with researchers working under the assumption that competition mitigates these behaviours and will eventually drive them from the market. The same is true of other–regarding behaviour (Schmidt, 2011). However, Chapter 4 highlights how such behaviours can exist in market places, and can have an impact on market outcomes. This conclusion is likely to be particularly true in de–centralised market places where bilateral face-to-face interactions are common place, although it is acknowledged that they are unlikely to play a role in centralised markets. In centralised markets humans do not interact *per se*, but instead interact with computers that facilitate transactions between market agents. Future research should consider the type of market, and the manner in which market agents interact, in order to determine if other–regarding preferences and discriminatory behaviour might play a role in the strategic settings being analysed.

As with all research, the research presented in this thesis has its limitations and the conclusions drawn from this thesis could be strengthened by the collection of additional data. For example, Chapter 2 provides a replication study that highlights how procedural differences in the laboratory may have consequences for behaviour. Additional experiments would add robustness to the inference made in that chapter. It would also be fruitful to supplement the findings from Chapter 2 with data from other replications of Fehr *et al.* (2007). For example, an analysis of the data collected from the on line laboratory of Charles Holt, the VECON Lab, which has a significant amount of data on the experimental game studied in Chapter 2, might provide additional insight into that chapter's conclusions.[1] This avenue of improvement is being actively pursued. This also raises the possibility that the experiments were statistically underpowered, or that the null results observed in each of the chapters is a result of poor experimental planning. Appendix D provides a retrospective power analysis, which shows that the experiments are powered at conventional levels.

In conclusion, this thesis highlights how discriminatory tastes can manifest themselves in social and other–regarding preferences. It explores how these preferences can have implications in both market and non–market settings, and uses structural modelling techniques to retrieve deep preference parameters. As far as the author of this thesis is aware, the research contained within this thesis presents the first estimates of social preference parameters obtained using real market interactions and the first to

[1] The VECON Lab can be found at the following address: http://veconlab.econ.virginia.edu/.

show that these preferences are influenced by ethnicity. As highlighted in the discussion of Al-Ubaydli & List (2016), structural modelling permits welfare and counterfactual analyses, and as such it is crucial for deepening economists' understanding of markets and market dynamics. It is hoped that the work contained in this thesis makes a contribution to that understanding.

Bibliography

Ahmed, A. M. (2007), 'Group identity, social distance and intergroup bias', *Journal of Economic Psychology* **28**(3), 324 – 337.

Ahmed, A. M. (2010*a*), 'Muslim discrimination: Evidence from two lost-letter experiments', *Journal of Applied Social Psychology* **40**(4), 888–898.

Ahmed, A. M. (2010*b*), 'What is in a surname? the role of ethnicity in economic decision making', *Applied Economics* **42**(21), 2715–2723.

Ahmed, A. M., Andersson, L., Hammarstedt, M. *et al.* (2009), 'Ethnic discrimination in the market place of small business transfers', *Economics Bulletin* **29**(4), 3050–3058.

Ahmed, A. M. & Hammarstedt, M. (2008), 'Discrimination in the rental housing market: A field experiment on the internet', *Journal of Urban Economics* **64**(2), 362 – 372.

Akerlof, G. A. (1982), 'Labor contracts as partial gift exchange', *Quarterly Journal of Economics* **97**(4), 543–569.

Akerlof, G. A. & Kranton, R. E. (2000), 'Economics and identity', *Quarterly Journal of Economics* **115**(3), 715–753.

Al-Ubaydli, O. & List, J. A. (2015), On the generalizability of experimental results in economics, *in* G. R. Frèchette & A. Schotter, eds, 'Handbook of Experimental Economic Methodology', Oxford University Press.

Al-Ubaydli, O. & List, J. A. (2016), Field experiments in markets. *National Bureau of Economic Research.*

Andreoni, J., Harbaugh, W. & Vesterlund, L. (2003), 'The carrot or the stick: Rewards, punishments, and cooperation', *American Economic Review* **93**(3), 893–902.

Apesteguia, J. & Ballester, A. M. (2016), Monotone stochastic choice models: The case of risk and time preferences.

Ayres, I., Vars, F. E. & Zakariya, N. (2005), 'To insure prejudice: Racial disparities in taxicab tipping', *The Yale Law Journal* **114**(7), 1613–1674.

Balafoutas, L., Beck, A., Kerschbamer, R. & Sutter, M. (2013), 'What drives taxi drivers? a field experiment on fraud in a market for credence goods', *The Review of Economic Studies* **80**(3), 876–891.

Bandiera, O., Barankay, I. & Rasul, I. (2005), 'Social preferences and the response to incentives: Evidence from personnel data', *Quarterly Journal of Economics* **120**(3), 917–962.

BBC (2015), 'Police 'sorry' over rochdale child sex abuse failures', The BBC, Website. Accessed September 2016.
URL: *http://www.bbc.co.uk/news/uk-england-manchester-31857066*

BBC (2016), 'Rochdale men jailed for 'appalling and vile' abuse', The BBC, Website. Accessed September 2016.
URL: *http://www.bbc.co.uk/news/uk-england-manchester-35982241*

Becker, G. S. (1971), 'The economics of discrimination', *University of Chicago Press Economics Books* .

Benson, P. L., Karabenick, S. A. & Lerner, R. M. (1976), 'Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help', *Journal of Experimental Social Psychology* **12**(5), 409–415.

Benz, M. & Meier, S. (2008), 'Do people behave in experiments as in the field? evidence from donations', *Experimental Economics* **11**(3), 268–281.

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

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

Bertrand, M. & Mullainathan, S. (2004), 'Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination', *American Economic Review* **94**(4), 991–1013.

Blanco, M., Engelmann, D. & Normann, H. T. (2011), 'A within-subject analysis of other-regarding preferences', *Games and Economic Behavior* **72**(2), 321–338.

Boffey, D. & Helm, T. (2016), 'Vote leave embroiled in race row over turkey security threat claims', The Guardian, Website. Accessed July 2016.
URL: *http://www.theguardian.com/politics/2016/may/21/vote-leave-prejudice-turkey-eu-security-threat*

Bolton, G. E., Katok, E. & Zwick, R. (1998), 'Dictator game giving: Rules of fairness versus acts of kindness', *International Journal of Game Theory* **27**(2), 269–299.

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

Booth, A. L., Leigh, A. & Varganova, E. (2012), 'Does ethnic discrimination vary across minority groups? Evidence from a field experiment', *Oxford Bulletin of Economics and Statistics* **74**(4), 547–573.

Bunyan, N. (2012), 'Rochdale grooming trial: Gang convicted for sex trafficking', The Telegraph, Website. Accessed September 2016.
URL: *http://www.telegraph.co.uk/news/uknews/crime/9239126/Rochdale-grooming-trial-gang-convicted-for-sex-trafficking.html*

Camerer, C. (2015), The promise and success of lab-field generalizability in experimental economics: A critical reply to levitt and list, *in* G. R. Frchette & A. Schotter, eds, 'Handbook of Experimental Economic Methodology', Oxford University Press.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. & Wu, H. (2016), Evaluating replicability of laboratory experiments in economics. *Science*.

Camerer, C. F. & Fehr, E. (2006), 'When does "economic man" dominate social behavior?', *Science* **311**(5757), 47–52.

Camerer, C. & Fehr, E. (2004), Measuring social norms and preferences using experimental games: A guide for social scientists, *in* J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr & H. Gintis, eds, 'Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies', Oxford University Press.

Carter, H. & Midlane, T. (2012), 'Rochdale violence raises fears over racial tensions in north-west england', The Guardian, Website. Accessed August 2014.

Castillo, M., Petrie, R., Torero, M. & Vesterlund, L. (2013), 'Gender differences in bargaining outcomes: A field experiment on discrimination', *Journal of Public Economics* **99**, 35–48.

Charness, G., Cobo-Reyes, R. & Jimnez, N. (2014), 'Identities, selection, and contributions in a public-goods game', *Games and Economic Behavior* **87**, 322–338.

Charness, G., Rigotti, L. & Rustichini, A. (2007), 'Individual behavior and group membership', *American Economic Review* **97**(4), 1340–1352.

Chen, R. & Chen, Y. (2011), 'The potential of social identity for equilibrium selection', *American Economic Review* **101**(6), 2562–89.

Chen, Y. & Li, S. X. (2009), 'Group identity and social preferences', *American Economic Review* **99**(1), 431–57.

Chen, Y., Li, S. X., Liu, T. X. & Shih, M. (2014), 'Which hat to wear? impact of natural identities on coordination and cooperation', *Games and Economic Behavior* **84**, 58–86.

Cohen, J. (1992), 'A power primer.', *Psychological bulletin* **112**(1), 155.

Cooper, D. & Kagel, J. H. (2009), 'Other regarding preferences: a selective survey of experimental results', *Handbook of experimental economics* **2**.

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

Cox, T. H., Lobel, S. A. & McLeod, P. L. (1991), 'Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task', *Academy of Management Journal* **34**(4), 827–847.

Crandall, C. S. & Sherman, J. W. (2016), On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology.*

Cuzick, J. (1985), 'A wilcoxon–type test for trend', *Statistics in medicine* **4**(4), 543–547.

DellaVigna, S. (2009), 'Psychology and economics: Evidence from the field', *Journal of Economic Literature* **47**(2), 315–72.

DellaVigna, S., John, A. L. & Malmendier., U. (2012), 'Testing for altruism and social pressure in charitable giving', *Quarterly Journal of Economics* **127**(1), 1–56.

Drouvelis, M. & Nosenzo, D. (2013), 'Group identity and leading-by-example', *Journal of Economic Psychology* **39**, 414–425.

Dufwenberg, M., Heidhues, P., Kirchsteiger, G., Riedel, F. & Sobel, J. (2011), 'Other-regarding preferences in general equilibrium', *The Review of Economic Studies* **78**(2), 613–639.

Eckel, C. C. & Grossman, P. J. (2005), 'Managing diversity by creating team identity', *Journal of Economic Behavior & Organization* **58**(3), 371–392.

Ely, J. C. & Valimaki, J. (2003), 'Bad reputation', *Quarterly Journal of Economics* **118**(3), 785–814.

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

Falk, A. (2007), 'Gift exchange in the field', *Econometrica* **75**(5), 1501–1511.

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

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

Fehr, E. & Gächter, S. (2000*a*), 'Cooperation and punishment in public goods experiments', *American Economic Review* **90**(4), 980–994.

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

Fehr, E., Gächter, S. & Kirchsteiger, G. (1997), 'Reciprocity as a contract enforcement device: Experimental evidence', *Econometrica* **65**(4), 833–860.

Fehr, E., Kirchsteiger, G. & Riedl, A. (1993), 'Does fairness prevent market clearing? an experimental investigation', *Quarterly Journal of Economics* **108**(2), 437–459.

Fehr, E., Kirchsteiger, G. & Riedl, A. (1998*a*), 'Gift exchange and reciprocity in competitive experimental markets', *European Economic Review* **42**(1), 1–34.

Fehr, E., Kirchsteiger, G. & Riedl, A. (1998*b*), 'When social norms overpower competition: Gift exchange in experimental labor markets', *Journal of Labor Economics* **16**(2), 324–351.

Fehr, E., Klein, A. & Schmidt, K. M. (2007), 'Fairness and contract design', *Econometrica* **75**(1), 121–154.

Fehr, E. & Schmidt, K. M. (1999), 'A theory of fairness, competition, and cooperation', *Quarterly Journal of Economics* **114**(3), 817–868.

Feltovich, N. (2003), 'Nonparametric tests of differences in medians: comparison of the Wilcoxon–Mann–Whitney and robust rank-order tests', *Experimental Economics* **6**(3), 273–297.

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

Fischbacher, U. (2007), 'z-tree: Zurich toolbox for ready-made economic experiments', *Experimental economics* **10**(2), 171–178.

Fowler, J. H. & Kam, C. D. (2007), 'Beyond the self: Social identity, altruism, and political participation', *Journal of Politics* **69**(3), 813–827.

Friedman, H. H., Herskovitz, P. J. & Pollack, S. (1994), The biasing effects of scale-checking styles on response to a likert scale, *in* 'Proceedings of the American statistical association annual conference: survey research methods', Vol. 792.

Gerber, A. S., Green, D. P. & Larimer, C. W. (2008), 'Social pressure and voter turnout: Evidence from a large-scale field experiment', *American Political Science Review* **102**, 33–48.

Gneezy, U. & List, J. A. (2006), 'Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments', *Econometrica* **74**(5), 1365–1384.

Gneezy, U., List, J. & Price, M. K. (2012), Toward an understanding of why people discriminate: Evidence from a series of natural field experiments, Working Paper 17855, National Bureau of Economic Research.

Goette, L., Huffman, D. & Meier, S. (2006), 'The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups', *American Economic Review* **96**(2), 212–216.

Goldin, C. & Rouse, C. (2000), 'Orchestrating impartiality: The impact of "blind" auditions on female musicians', *American Economic Review* **90**(4), 715–741.

Grosskopf, B. & Sarin, R. (2010), 'Is reputation good or bad? An experiment', *American Economic Review* **100**(5), 2187–2204.

Guala, F. & Filippin, A. (2015), The effect of group identity on distributive choice: Social preference or heuristic? *The Economic Journal*.

Guala, F., Mitten, L. & Ploner, M. (2013), 'Group membership, team preferences, and expectations', *Journal of Economic Behavior & Organization* **86**, 183–190.

Haley, K. J. & Fessler, D. M. (2005), 'Nobody's watching?: Subtle cues affect generosity in an anonymous economic game', *Evolution and Human behavior* **26**(3), 245–256.

Harrison, G. W. & List, J. A. (2004), 'Field experiments', *Journal of Economic Literature* **42**(4), 1009–1055.

Heckman, J. J. (1998), 'Detecting discrimination', *Journal of Economic Perspectives* **12**(2), 101–116.

Hinsliff, G. (2016), 'The tories attempts to smear sadiq khan are repellen', The Guardian, Website. Accessed July 2016.
URL: *https://www.theguardian.com/commentisfree/2016/apr/22/tories-attempt-smear-sadiq-khan-repellent*

Hoenig, J. M. & Heisey, D. M. (2001), 'The abuse of power', *The American Statistician* **55**(1), 19–24.

Hoffman, E., McCabe, K., Shachat, K. & Smith, V. (1994), 'Preferences, property rights, and anonymity in bargaining games', *Games and Economic Behavior* **7**(3), 346–380.

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

Holm, H. J. (2000), 'Gender-based focal points', *Games and Economic Behavior* **32**(2), 292–314.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**(2), 65–70.

Ioannou, C. A., Qi, S. & Rustichini, A. (2015), 'Group payoffs as public signals', *Journal of Economic Psychology* **48**, 89–105.

Jackson, C. K. & Schneider, H. S. (2011), 'Do social connections reduce moral hazard? evidence from the new york city taxi industry', *American Economic Journal: Applied Economics* **3**(3), 244–67.

Kaas, L. & Manger, C. (2012), 'Ethnic discrimination in germany's labour market: A field experiment', *German Economic Review* **13**(1), 1–20.

Koch, A. K. & Normann, H.-T. (2008), 'Giving in dictator games: Regard for others or regard by others?', *Southern Economic Journal* **75**(1), 223–231.

Kube, S., Maréchal, M. A. & Puppe, C. (2012), 'The currency of reciprocity: Gift exchange in the workplace', *American Economic Review* **102**(4), 1644–1662.

Landry, C. E., Lange, A., List, J. A., Price, M. K. & Rupp, N. G. (2006), 'Toward an understanding of the economics of charity: Evidence from a field experiment', *Quarterly Journal of Economics* **121**(2), 747–782.

Levitt, S. D. (2004), 'Testing theories of discrimination: Evidence from the weakest link', *Journal of Law and Economics* **47**(2), 431–452.

Levitt, S. D. & List, J. A. (2007), 'What do laboratory experiments measuring social preferences reveal about the real world?', *Journal of Economic Perspectives* **21**(2), 153–174.

List, J. A. (2004), 'The nature and extent of discrimination in the marketplace: Evidence from the field', *Quarterly Journal of Economics* **119**(1), 49–89.

List, J. A. (2006), 'The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions', *Journal of Political Economy* **114**(1), 1–37.

List, J. A., Shaikh, A. M. & Xu, Y. (2016), Multiple hypothesis testing in experimental economics.

Malmendier, U., te Velde, V. L. & Weber, R. A. (2014), 'Rethinking reciprocity', *Annual Review of Economics* **6**(1), 849–874.

Mas, A. & Moretti, E. (2009), 'Peers at work', *American Economic Review* **99**(1), 112–45.

Mateos, P., Webber, R. & Longley, P. A. (2007), The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names. University College London, Working Paper.

Milibank, D. (2015), 'Donald trump is a bigot and a racist', The Washington Post, Website. Accessed January 2016.

Mitra, A. & Ray, D. (2014), 'Implications of economic theory of conflict: Hindu–muslim violence in India', *Journal of Political Economy* **122**(4), 719–765.

Mobius, M. M. & Rosenblat, T. S. (2006), 'Why beauty matters', *American Economic Review* **96**(1), 222–235.

Mujcic, R. & Frijters, P. (2013), Still not allowed on the bus: It matters if you're black or white! IZA Discussion Paper.

Öhman, A. (1986), 'Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion', *Psychophysiology* **23**(2), 123–145.

Oreopoulos, P. (2011), 'Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes', *American Economic Journal: Economic Policy* **3**(4), 148–71.

Page, B. (2009), 'British attitudes to immigration in the 21st century', Migration Policy Institute.

Phelps, E. S. (1972), 'The statistical theory of racism and sexism', *American Economic Review* **62**(4), 659–661.

Read, D., Loewenstein, G. & Rabin, M. (1999), 'Choice bracketing', *Journal of Risk and Uncertainty* **19**(1), 171–197.

Riach, P. A. & Rich, J. (2002), 'Field experiments of discrimination in the market place', *The Economic Journal* **112**(483), 480–518.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. (1991), 'Bargaining and market behavior in jerusalem, ljubljana, pittsburgh, and tokyo: An experimental study', *American Economic Review* **81**(5), 1068–1095.

Ruffle, B. J. & Shtudiner, Z. (2015), 'Are good-looking people more employable?', *Management Science* **61**(8), 1760–1776.

Ruffle, B. J. & Sosis, R. (2006), 'Cooperation and the in-group-out-group bias: A field test on israeli kibbutz members and city residents', *Journal of Economic Behavior & Organisation* **60**(2), 147–163.

Schmidt, K. M. (2011), 'Social preferences and competition', *Journal of Money, credit and Banking* **43**(1), 207–231.

Schneider, H. S. (2012), 'Agency problems and reputation in expert services: Evidence from auto repair', *Journal of Industrial Economics* **60**(3), 406–433.

Schupp, H. T., Öhman, A., Junghöfer, M., Weike, A. I., Stockburger, J. & Hamm, A. O. (2004), 'The facilitated processing of threatening faces: An ERP analysis', *Emotion* **4**(2), 189.

Schwab, S. (1986), 'Is statistical discrimination efficient?', *American Economic Review* **76**(1), 228–234.

Smith, V. L. (1994), 'Economics in the laboratory', *Journal of Economic Perspectives* **8**(1), 113–131.

Stoop, J. (2014), 'From the lab to the field: envelopes, dictators and manners', *Experimental Economics* **17**(2), 304–313.

Stoop, J., Noussair, C. N. & Van Soest, D. (2012), 'From the lab to the field: Cooperation among fishermen', *Journal of Political Economy* **120**(6), 1027–1056.

Syal, R. & Topping, A. (2014), 'Rochdale taxi firm admits providing white drivers on request', The Guardian, Website. Accessed November 2014.

Tajfel, H., Billig, M., G. Bundy, R. P. & Flament, C. (1971), 'Social categorization and intergroup behaviour', *European Journal of Social Psychology* **1**(2), 149–178.

Tajfel, H. & Turner, J. (1986), The social identity theory of intergroup behavior, *in* S. Worchel & W. Austin, eds, 'The Psychology of Intergroup Relations', Chicago: NelsonHall, pp. 7–24.

Thompson, D. (2016), 'Who are donald trump's supporters, really?', The Atlantic, Website. Accessed July 2016.
URL: *http://www.theatlantic.com/politics/archive/2016/03/who-are-donald-trumps-supporters-really/471714/*

Townsend, M. (2012), 'Voters more likely to back an anti–muslim party than reject it', The Guardian, website. Accessed January 2016.

Tsutsui, K. & Zizzo, D. J. (2014), 'Group status, minorities and trust', *Experimental Economics* **17**(2), 215–244.

Turner, J. C. (1975), 'Social comparison and social identity: Some prospects for intergroup behaviour', *European Journal of Social Psychology* **5**(1), 1–34.

Turner, J. C. (1978), Social categorization and social discrimination in the minimal group paradigm., *in* H. Tajfel, ed., 'Differentiation between social groups: Studies in the social psychology of intergroup relations', London: Academic Press, pp. 101–140.

van Der Mewe, G. W. & Burns, J. (2008), 'What's in a name? Racial identity and altruism in post-apartheid South Africa', *South African Journal of Economics* **76**(2), 266–275.

Weng, L.-J. & Cheng, C.-P. (2000), 'Effects of response order on likert-type scales', *Educational and psychological measurement* **60**(6), 908–924.

Whitt, S. & Wilson, R. K. (2007), 'The dictator game, fairness and ethnicity in postwar Bosnia', *American Journal of Political Science* **51**(3), 655–668.

Winking, J. & Mizer, N. (2013), 'Natural-field dictator game shows no altruistic giving', *Evolution and Human Behavior* **34**(4), 288–293.

Wright, O. (2016), 'Nigel farage accused of deploying nazi-style propaganda as remain crash poster unveiling with rival vans', The Independent, Website. Accessed July 2016.
URL: *http://www.independent.co.uk/news/uk/politics/nigel-farage-brexit-poster-vans-eu-referendum-london-remain-breaking-point-a7085396.html*

Xiao, E. & Houser, D. (2005), 'Emotion expression in human punishment behavior', *Proceedings of the National Academy of Sciences of the United States of America* **102**(20), 7398–7401.

Yamagishi, T. & Kiyonari, T. (2000), 'The group as the container of generalized reciprocity', *Social Psychology Quarterly* **63**(2), 116–132.

Zizzo, D. J. (2010), 'Experimenter demand effects in economic experiments', *Experimental Economics* **13**(1), 75–98.

Zizzo, D. J. (2012), Inducing natural group identity: A RDP analysis. University of East Anglia Discussion Paper, 12-03.

Appendix A

Appendix to Chapter 2

This appendix provides the experimental material for Chapter 2, including instructions and screenshots of the zTree interface.

A.1 Experimental Appendix

NB: The appearance of these instructions differs slightly to those used in the experiment due to differences in document formatting.

Introduction

You are about to take part in an experiment in decision-making. Various institutions have provided the funds for this research. The total amount of money you earn depends upon the decisions you make and on the decisions that other people make. Throughout the experiment we will speak in terms of tokens and not pounds. At the end of the experiment you will be paid in cash based on the following exchange rate:

1 token = 4p

You will be given an initial endowment of 100 tokens. In addition to this initial endowment you have already earned a 2.50 show up fee. At the end of the experiment, everyone will be paid in private, with no obligation to tell others how much you earn. It is also important to note that any decisions you make will remain anonymous throughout the session and once the experimental session is complete.

It is important that you remain silent and do not look at other people's work. If you have any questions, or need assistance of any kind, please raise your hand and an experimenter will come to you. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your following of these rules.We will first jointly go over the instructions. Once the instructions have been read through, you will have time to ask any clarifying questions by raising your hand. Do not touch the computers until you are instructed to.

Instructions

In this experiment you will take part in a task. You will be randomly assigned to a Role, the role of Principal or the role of Agent. There are 20 people in this experiment and exactly half of you are assigned as Principals, the other half as Agents. There are therefore 10 Agents and 10 Principals. If you are assigned as a Principal you will remain a Principal for the entire session; if you are assigned as an Agent you will remain an Agent for the entire session. We will now go through the instructions for the task.

The task involves 10 Periods. In each Period you will be matched to one other person. If you are a Principal you will always be matched to an Agent. You will be matched to every Agent exactly once, in a random order. If you are an Agent you will always be matched to a Principal. You will be matched to every Principal exactly once, in a random order. In each period an information box is displayed which reminds you of your Role and the Role of the person you are matched to.

Each of the 10 Periods consists of 3 sequential Stages. In each stage that you are required to make a decision, you have a maximum of 60 seconds to make it. The time limit is always displayed in the top right hand corner of your screen. If time runs out a message will be displayed encouraging you to reach a decision. It is important to understand that the experiment cannot continue to the next stage until everyone has made a decision. We will now explain each stage in detail.

In stage 1 the Principal makes a decision. The Principal must first decide whether or not to offer a contract to the Agent. If no contract is offered both get 0 tokens in that Period and both must wait until the next Period. If the Principal makes an offer, they must specify the following 3 components of a contract:

1. a fixed wage offer in tokens to be paid to the Agent, w. This choice can be any whole number between 0 and the total amount of tokens the Principal has. If the Agent accepts the contract they will receive this wage for certain.

2. demand an effort level from the Agent, e^*. The Agent has 10 effort levels to choose from. One must be specified from between 1 and 10. The actual effort choice the Agent makes can be different from this value.

3. an announced bonus, in tokens, which could be given to the Agent as an additional grant in Stage 3, b^*. This can take any whole number from 0 to 100. A screen shot of Stage 1 is provided below:

   The screen consists of a number of boxes and tables:

   (a) The Information Box.

   (b) The Agents Effort-Cost Table.

   Effort is costly to the Agent. This table displays the costs associated with each effort level the Agent could choose. As an example, an effort choice of 1 has a cost of 0 tokens associated with it whilst a choice of 10 has a cost of 20 tokens.3. The Decision Box. The Principal makes their choices in this box. They initially must decide whether or not to make an offer, pressing continue to make their choice. If they decide to make an offer, the box will change and look as follows:

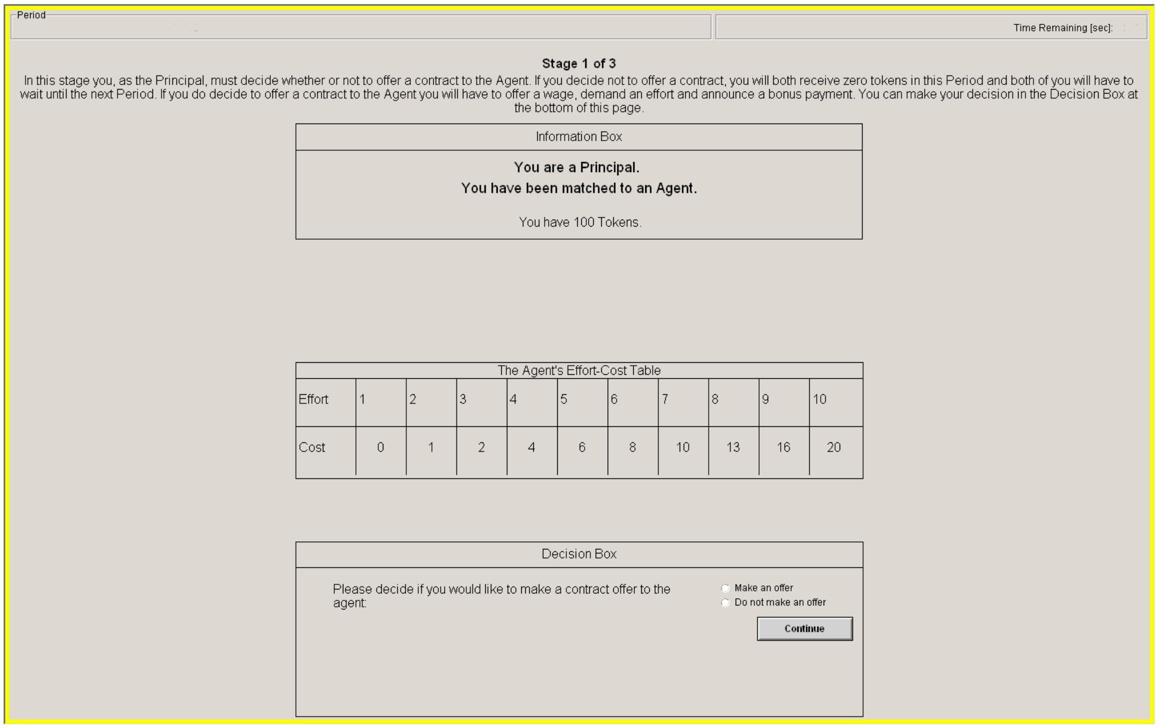


Figure A.1: Stage 1

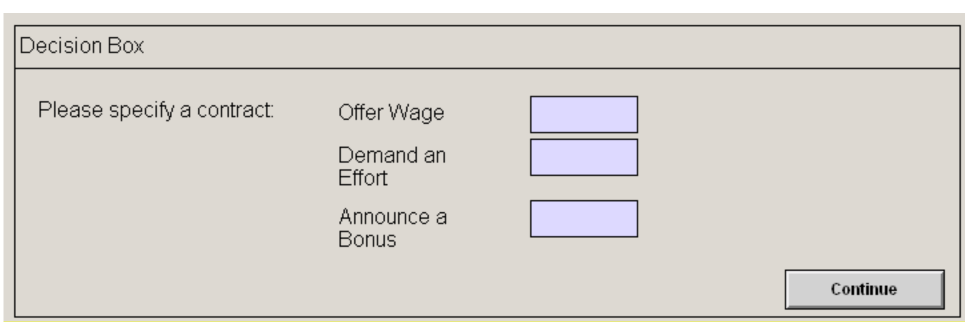


Figure A.2: Box 1

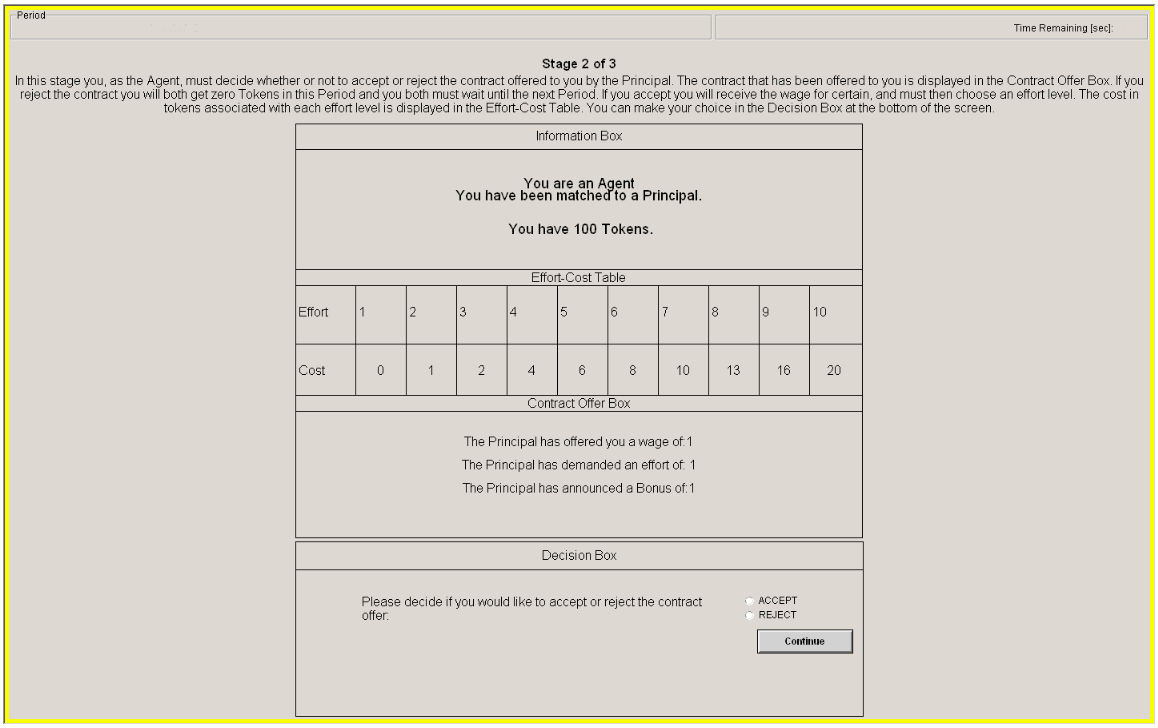


Figure A.3: Stage 2

Once the Principal has made a decision they must click continue; the stage ends and moves to Stage 2.

In Stage 2 The Agent must make a decision in Stage 2. If a contract is offered, the Agent can first either accept or reject it. If the contract is rejected, both the Principal and Agent get 0 tokens that Period and must wait until the next Period. If the Agent accepts the contract, they receive the wage offer w for certain. The Agent must then make a decision:

- choose an effort level e.

The effort choice e is allowed to differ from the effort demanded by the Principal e^*. Effort is costly and for the effort choice e, the Agent incurs a cost in tokens. The Agent can choose an effort from 1 to 10, and the costs for each effort level are shown in the Agent's Effort-Cost table: A screen shot of Stage 2 is provided below:

The screen consists of a number of boxes and tables:

1. Information Box

2. The Agents Effort-Cost Table.

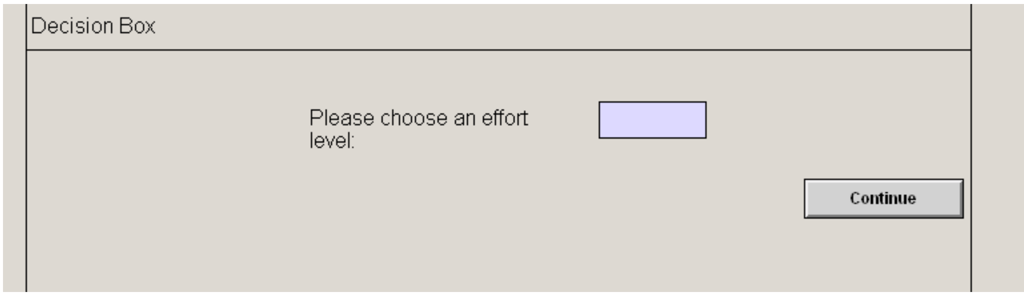


Figure A.4: Box 2

3. The Contract Offer Box. This box outlines the contract offered by the Principal to the Agent, if one has been offered. If one hasn't been offered, the Agent will be informed.

4. The Decision Box. The Agent makes their choice in this box. They must first decide whether to accept or reject, clicking the Continue button to make their choice. If they reject, both must wait until the next period. If they accept, the decision box changes as shown below

Once the Agent has made a choice they must click continue; the stage ends and moves to Stage 3.

The Principal makes a decision in Stage 3. If the Agent rejects the contract, the Principal is informed. If the Agent accepts the contract and chooses an effort level, the Principal must then decide on an actual bonus payment, b. This bonus payment can differ from b^* the announced bonus payment made in Stage 1. The actual bonus payment can be between 0 and 100.

Stage 3 looks as follows:

The Profits received by the Principal and Agent are calculated at the end of each Period. If the Principal does not offer a contract or the Agent rejects the offer both the Principal and the Agent get 0 Tokens that Period. If a contract is offered and accepted, the Profits in each period are calculated as follows:

The Principal : $\pi_p = 10e - w - b$ (The effort choice of the Agent multiplied by 10, minus the wage and minus the actual bonus payment to the Agent), the Agent: $\pi_A = w$-effortcost$+b$

(The wage minus the Agent's effort-cost plus the actual bonus payment from the Principal) Your Total Profit is the sum of all your Profits from previous Periods plus your initial endowment. It should be noted that the number of tokens the Principal

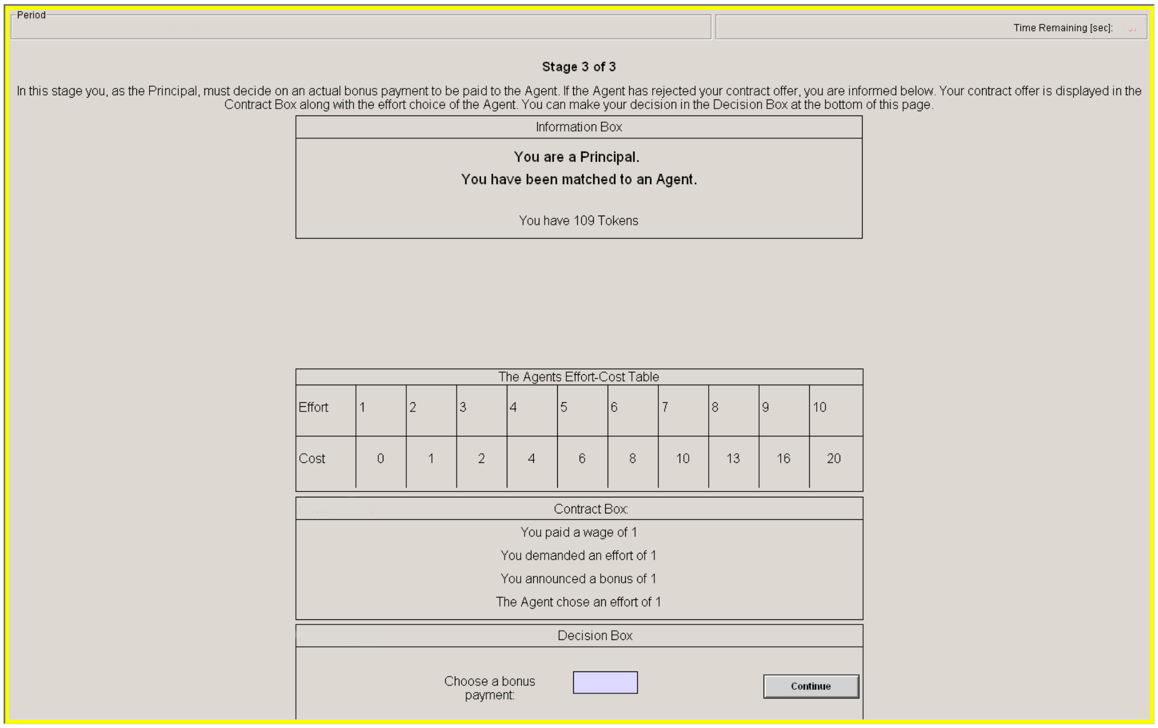


Figure A.5: Stage 3

has, as shown in the information box, updates between Stages. In Stage 3, if the Agent has accepted the contract in Stage 2 and selected an effort choice, the tokens that the Principal has earned so far this period are added to their tokens in the information box (ten times the Agent effort choice from Stage 2 is added whilst the wage that was paid in Stage 1 is subtracted). The actual bonus payment is subtracted from the Principal's tokens at the end of Stage 3.

At the end of each period you will be shown the outcome from that Period and then rematched to another participant and the next period will begin.

Practice Exercises

NB: There were three practice exercises given to subjects, for the sake of space, only one is outlined below.

We will now go through two practice exercises to ensure your understanding of the Task. Assume for these questions that both the Principal and the Agent are initially endowed with 100 tokens. Please refer to the instructions above to aid you in answering the questions. Once you have answered all the questions, please raise your hand and an experimenter will come and check your answers.

In Period 1, suppose the Principal offers the following contract in stage 1: w=10,e^*=10 and b^*=10. Suppose the Agent accepts and chooses the following effort level in stage 2:$e = 3$ and the Principal pays the following bonus in stage 3:$b = 5$

- What is the Principal's Profit at the end of this Period?
- What are the Principal's Total Profit at the end of this Period?
- What is the cost to the Agent of the effort choice (you may wish to look at the Agent's effort-cost table)?
- What is the Agent's Profit at the end of this Period?
- What are the Agent's Total Profit at the end of this Period?
- If the Agent had rejected the contract offer in Period 1, stage 2 what would the Profits for the Principal and Agent be?

If you have any further questions now the exercise has finished, please raise your hand and an experimenter will come to you.

You can now begin the Task by pressing Continue on your screen.

A.1.1 Instructions for the identity inducement procedure

Introduction

You are about to take part in an experiment in decision-making. Various institutions have provided the funds for this research. The total amount of money you earn depends upon the decisions you make and on the decisions that other people make. Throughout the experiment we will speak in terms of tokens and not pounds. At the end of the experiment you will be paid in cash based on the following exchange rate:

1 token = 4p

You will be given an initial endowment of 100 tokens. In addition to this initial endowment you have already earned a 2.50 show up fee. At the end of the experiment, everyone will be paid in private, with no obligation to tell others how much you earn. It is also important to note that any decisions you make will remain anonymous throughout the session and once the experimental session is complete.

It is important that you remain silent and do not look at other people's work. If you have any questions, or need assistance of any kind, please raise your hand and an experimenter will come to you. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your following of these rules.We will first jointly go over the instructions. Once the instructions have been read through, you will have time to ask any clarifying questions by raising your hand. Do not touch the computers until you are instructed to.

Instructions

The experiment is split into 2 parts, Part 1 and Part 2. In each part you will be asked to make one or more decisions and will have a chance to earn money. The amount of money you will earn in each part of the experiment will depend on your decisions and may depend on other participants' decisions. The total amount you will earn from the experiment will be the sum of the earnings you make in the two parts of the experiment. You will be informed about your earnings from Part 1 at the end of the session. Therefore in Part 2 you will make your decisions without knowing any outcome from Part 1. We will now jointly go over the instructions for Part 1. You will receive the instructions for Part 2 once everyone has completed Part 1.

Part 1

In PART 1 of the experiment you and the other participants will be assigned to a group. There are twenty participants in this experiment. Ten participants will be randomly assigned to the RED group, and ten other participants will be assigned to the GREEN group. You will not learn the identity of the other participants in your group, during or after today's session.

Once every participant has been randomly assigned to a group, you will be shown a screen with six paintings that belong to two artists: Artist A and Artist B. You will be given 180 seconds (3 minutes) to inspect the paintings. During these 180 seconds, you can use a group chat program to discuss the paintings with the other members in your own group. Messages will be shared only among the members of your own group. You will not be able to see the messages exchanged among the members of the other group. People in the other group will not see the messages from your group. Once 180 seconds has elapsed, your screen will change. You will then be shown four more paintings and you will need, for each painting, to select the artist who you think made the painting. You will be given 180 seconds (3 minutes) for this task. During these 180 seconds, you can continue using the group chat program to get help from or offer help to the other members in your own group. Messages will be shared only among the members of your own group. For each correct answer, you will be rewarded with 20 tokens.

When you use the chat program, you can type whatever you want in the lower box of the chat program, except for the following restrictions.

Restrictions on messages

1. You must not identify yourself or send any information that could be used to identify you (for example, your name, contact details or seat in the room)

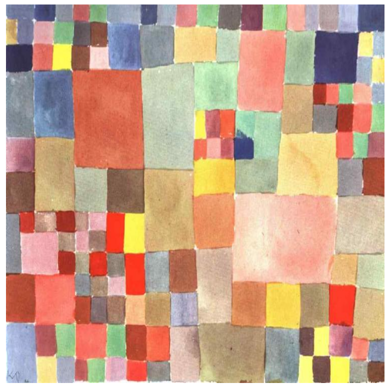
2. You must not make any threats, insults or use any obscene or offensive language. If you violate these rules your payment will be forfeited.

You will be paid in private and in cash at the end of the experiment. Please raise your hand if you have any questions.

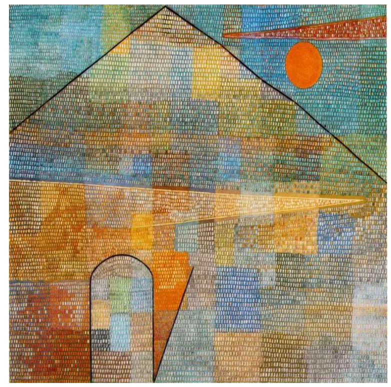
A.2 Pictures used in the identity study

Figure A.6 shows the three Paul Klee and three Wassily Kandinsky paintings that subjects had to consider privately in the identity inducement procedure. Figure A.7 shows the two Paul Klee and two Wassily Kandinsky paintings that subjects had to identify whilst being able to chat to their group members. The artist, title, and year the pictures were painted, are given below each picture.

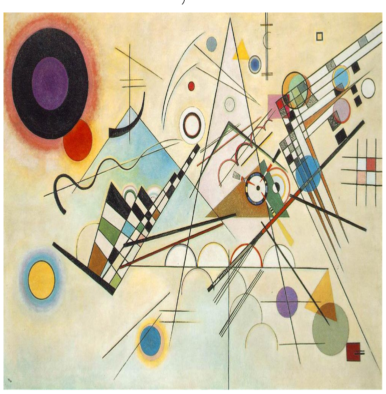
(a) Paul Klee: *Castle and Sun*, 1928



(b) Paul Klee: *Flora on Sand*, 1927

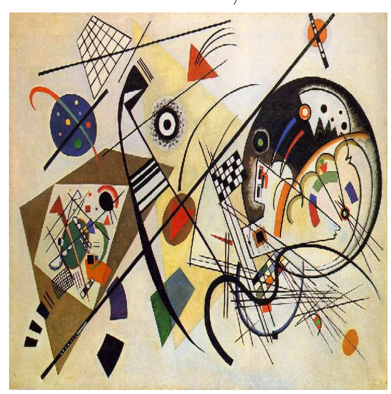


(c) Paul Klee: *Ad Parnassum*, 1932



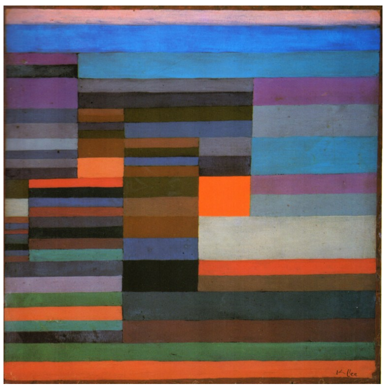
(d) Wassily Kandinsky: *Composition VIII*, 1923

(e) Wassily Kandinsky: *Composition IV*, 1911

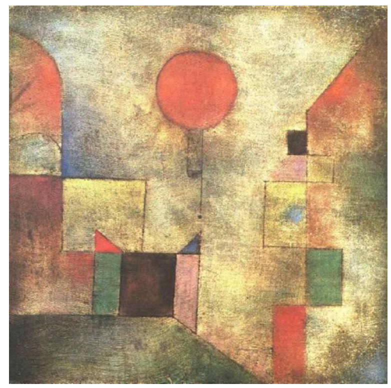


(f) Wassily Kandinsky: *Transverse Line*, 1923

Figure A.6: The Six Paintings Subjects were Required to Consider Privately in the *Identity* Study.

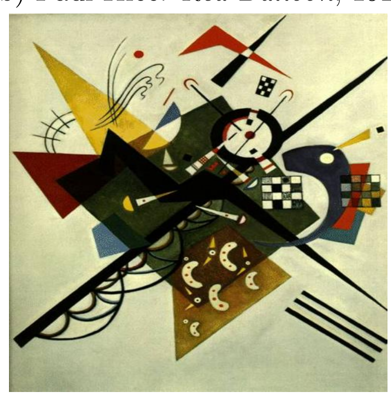


(a) Paul Klee: *Fire in the Evening*, 1929



(b) Paul Klee: *Red Balloon*, 1922

(c) Wassily Kandinsky: *Several Circles*, 1926



(d) Wassily Kandinsky: *On White II*, 1923

Figure A.7: The Four Paintings Subjects were Asked to Identify in the *Identity* Study.

Appendix B

Appendix to Chapter 3

This appendix provides supplementary material for Chapter 3. Section B.1 provides the experimental materials used in the data collection, including instructions and decision sheets. Section B.2 provides survey responses not included in the main text. Section B.3 provides an in-depth derivation of the likelihood function used in Section 3.5.

B.1 Experimental Appendix

B.1.1 Experimental instructions - Part 1

Instructions (Part 1)

In Part 1 you will be asked to make three choices. These choices will be referred to as **Choice 1**, **Choice 2** and **Choice 3**. In each **Choice**, you will be randomly **matched** to two people chosen from a list compiled from the Edited Electoral Register for the Rochdale area. These two people will be called **Person A** and **Person B**. **Person A** and **Person B** will not be required to make a decision.

In each **Choice**, the surnames of **Person A** and **Person B** will be revealed. No other information about these people can be given. Even if the surnames are identical between **Choice**s, the people will be different.

In each **Choice**, you will be asked how to split **10 pounds** between **Person A** and **Person B**. You can allocate the 10 pounds however you like, as long as the allocation adds up to 10 pounds and the amount given to each person is in whole pounds. **You cannot allocate money to yourself**.

Once the study is complete, one of these three **Choices** will be chosen at random, and **Person A** and **Person B** from that **Choice** will receive payment in cash.

The **Person A** and **Person B** that are chosen to receive payment, will not know they have taken part in this experiment until they receive a letter with the amount of money you have allocated them. The amount you decide to send will be placed in an envelope and will be delivered through **Person A's** and **Person B's** door. **Person A** and **Person B** will not learn any information about you.

You will not earn any money in Part 1.

When a **Choice** is chosen, the amount of money **Person A** and **B** will earn in pounds will be:

The amount that you allocate them.

To make sure you fully understand what is asked of you, please complete the three practice questions below. Once you have finished, please let the instructor know so that they can check your answers:

1. How much will **you** earn in Part 1?___________________________
2. If you allocate **Person A** 2 pounds:
   (a) How much will **Person A** earn?_________________
   (b) How much will **Person B** earn?________________
3. If you allocate **Person A** 8 pounds:
   (a) How much will **Person A** earn?_________________
   (b) How much will **Person B** earn?________________

B.1.2 Other–other decision sheets

Part 1 Answer Sheet
Choice 1

You must allocate £10 between Person A and Person B.

Person A's surname is__________________________

Person B's surname is__________________________

Write the amount of pounds you would like to allocate to each Person in the boxes below. The amounts allocated must add up to 10.

Amount Allocated to Person A	Amount Allocated to Person B

Once you have made your decision please turn over.

B.1.3 Experimental instructions - Part 2

Instructions (Part 2)

In Part 2 you will be given £10. You will then be **matched** to a person randomly chosen from a list compiled from the Edited Electoral Register for the Rochdale area. This person will be called **your match**. **Your match** is **not** the same person as anyone you allocated money to in Part 1. Before you make your decision I have to tell you the surname of **your match**.

Their surname is _________________. I can't reveal any other information about **your match**.

You will be required to make a single decision. **Your match** will not be required to make a decision.

You will be asked how many of the 10 pounds that you have been given you would like to **send** to **your match**. You can make any choice between 0 and 10. Your choice must be in whole pounds and not pence. The amount you decide to send will be placed in an envelope and will be delivered through **your match's** door with an accompanying letter. **Your match** will not know they have taken part in this experiment until they receive a letter with the amount of money you decide to send them. **Your match** will not learn your surname or any information about you.

The amount of money **you** earn in Part 2 in pounds will be :

10 minus the number of pounds you send.

The amount of money **your match** will earn in pounds will be:

The amount that you send them.

To make sure you fully understand what is asked of you, please complete the three practice questions below. Once you have finished, please let the instructor know so that they can check your answers:

1. If you send £0 how much will:
   (a) **you** earn in Part 2?___________________________
   (b) **your match** earn?_________________
2. If you send £5 how much will:
   (a) **you** earn in Part 2?___________________________
   (b) **your match** earn?_________________
3. If you send £10 how much will:
   (a) **you** earn in Part 2?___________________________
   (b) **your match** earn?_________________
4. Does **your match** know they are taking part?__________

B.1.4 Dictator game decision sheet

Part 2 Answer Sheet

You have been given **£10.**

Your match's surname is__________________________

Write the amount of pounds you would like to **send** to **your match** in the box below.

Once you have made your decision please turn over.

B.2 Statistical Appendix

B.2.1 Survey and responses

Variable	Observations	Mean	Std. Dev.
Male, 1 if yes	122	0.443	0.499
Married, 1 if yes	121	0.372	0.634
No. Children	113	0.628	0.485
Employed$^{\diamond}$	119	0.429	0.696
Income*	100	0.36	0.612
Education†	117	0.906	0.83
Housing benefit, 1 if yes	121	0.678	0.469
Years Living in Rochdale	120	26.142	16.353
Beliefs**	103	3.738	3.106

Note: Observations differ due to missing entries.
$^{\diamond}$0 if unemployed, 1 if employed, 2 if retired.
*0 if income < £10000, 1 if £10000 < income ≤ £20,0000, 2 if £20000 < income ≤ £30,0000 .
†0 if no qualifications, 1 if GCSE level, 2 if A–level, 3 if Degree, 4 if postgraduate.
** Subjects' belief about the experimenter's expectation of their behaviour in the Dictator Game.

B.3 Structural Appendix

B.3.1 Constructing the likelihood function

Fixing the parameters α, θ, a, b, for each $x \in \{0, 1, ..., 10\}$ we can determine the critical values of ϵ where the subjects utility maximising choice changes, ϵ_x. This is because the utility maximising choice of x, x^*, varies with ϵ. The dictator will choose to send an amount x over $x + 1$ up until,

$$u(x; \alpha, \theta, a, b, \epsilon_x) = u(x + 1; \alpha, \theta, a, b, \epsilon_x). \tag{B.1}$$

Rearranging for ϵ_x gives

$$\epsilon_x = \frac{(s + \omega - x)^{\alpha} - (s + \omega - x - 1)^{\alpha}}{(x + 1)^{\alpha} - x^{\alpha}} - \theta(1 + ae + bm) \tag{B.2}$$

Dividing through by σ gives,

$$\frac{\epsilon_x}{\sigma} = \frac{1}{\sigma}\left(\frac{(s+\omega-x)^{\alpha}-(s+\omega-x-1)^{\alpha}}{(x+1)^{\alpha}-x^{\alpha}} - \theta(1+ae+bm)\right) \tag{B.3}$$

When $\epsilon \in (\epsilon_{x-1}, \epsilon_x)$, $x^* = x$. The probability of choosing $x^* = x$ can then be determined from the cumulative distribution function of the error term. Where $f(z)$ is the density function, and $F(z)$ the cumulative distribution, the probability that the dictator chooses $x^* = 0$ is the probability that $\epsilon \in (-\infty, \epsilon_0)$, or

$$\Pr[x^* = 0|\alpha, \theta, a, b, \sigma] = \int_{-\infty}^{\epsilon_0} f(z)dz = F(\epsilon_0). \tag{B.4}$$

The probability the dictator chooses $x^* = x \in \{1, 2, ..., 9\}$ is

$$\Pr[x^* = x|\alpha, \theta, a, b, \sigma] = \int_{\epsilon_{x-1}}^{\epsilon_x} f(z)dz = F(\epsilon_x) - F(\epsilon_{x-1}), \tag{B.5}$$

and the probability of choosing $x^* = 10$ is

$$\Pr[x^* = 10|\alpha, \theta, a, b, \sigma] = \int_{\epsilon_9}^{\infty} f(z)dz = 1 - F(\epsilon_9).$$

The likelihood function, as we have $k = 122$ observations, is therefore

$$L(\alpha, \theta, a, b, \sigma) = \prod_{k=1}^{122} Pr[x_k = x; \alpha, \theta, a, b, \sigma]. \tag{B.6}$$

Appendix C

Appendix to Chapter 4

This appendix provides supplementary material for Chapter 4. Section C.1 provides the experimental materials used in the data collection, including job advertisements, experimental script sheets, decision sheets and surveys. Section C.2 presents additional derivations, summary statistics and robustness checks.

C.1 Experimental Appendix

C.1.1 Job advertisement

UNIVERSITY OF EXETER

48 Research Assistants Needed – No Previous Experience, Qualifications or Knowledge Required. All applications welcome!

- 48 Positions
- You will be paid £7.50 per hour.
- Location: Manchester
- **No previous experience or specialist knowledge required**
- Help out with ground breaking Economic research whilst getting paid!

We are looking for 48 individuals to help us conduct some economic research. You will begin by receiving training, and then be asked to complete a task. This task is very simple. The work is not recurring, and is a onetime offer from the researchers. Researchers from University of Exeter are conducting this research.

Due to the nature of the research, the exact task will only be revealed to successful applicants. However, the task will involve travelling on foot for short distances. Some knowledge of Manchester City Centre is a definite bonus. **It cannot be stressed enough that no prior experience, knowledge, or qualifications in any academic discipline are required. We welcome, and encourage, all applications.**

Applicants should be trustworthy, have the ability to follow instructions diligently, be able to read, and write in English and have good English speaking skills. We strongly encourage applications from all types of people, from all different walks of life.

Applicants should submit a short CV, in word, PDF, or in the body of an email to the email address provided below. You should also submit a passport sized photo. **Please also submit contact details, including a phone number and email address.** Successful applicants will be invited to attend a short training session in Manchester at a later date. By submitting an application, you agree to have your application reviewed by a specialist panel. If you are successful, the researchers will require this picture before you can take part in the task.

This research study has been reviewed by the Humanities & Social Sciences Ethical Review Committee at the University of Exeter. Applicants must have the right to work in the UK. Proof of this will be required if you are successful.

Email:

C.1.2 Experimental script sheet

Script Sheet		
Step	**Event**	**Speak / Action**
1	Approach the Taxi at the front of the rank.	
2	State Destination to Driver	To Driver: **I would like to go to destination X**
3	Enter Taxi	To Driver: **I don't take taxis very often**
4	Once the meter reaches £3 speak:	To Driver: **I'm sorry, I only have £4! Could you take me to my destination for that amount?**
4a	The driver gets irate	**Say nothing.**
4b	The Driver Offers to take you to a cash point	To Driver: **I don't have my bank card. (Repeat if necessary)**
5	The Driver tells you he will not take you	To Driver: **OK. Please will you take me as far as you can.**
6	The Driver Stops the Taxi	**Pay the driver**
		To Driver: **Please can I have a receipt?**
6a	**Important Step**	**Complete Record Sheet - NOTING DOWN THE METER READING**

C.1.3 Experimental sheet

Tester ID: Ride ID:

Taxi Rank: **RANK** Destination: **Destination**

	Questions about the Driver	(Tick where appropriate)			
1	Race / Ethnicity	White-British	☐	Mixed Race	☐
		East Asian (Chinese)	☐	Black	☐
		South Asian (Indian / Pakistani)	☐	White-Other	☐
2	Gender	Male	☐		
		Female	☐		
3	Age				
4	Raining	Yes	☐		
		No	☐		
5	Traffic (1= Not busy 10=Busy)				
6	Driver tried to have a conversation	Yes	☐		
		No	☐		
7	Driver Offered to take you to a cash point	Yes	☐		
		No	☐		
8	Driver Completed the Journey	Yes	☐		
		No	☐		
9	**Meter Reading when you left the taxi**				
10	Did the driver give you a receipt?	Yes	☐		
		No	☐		

C.1.4 Ex-post picture rating experimental instructions

Picture Rating Instructions

- You will be shown 11 pictures of different peoples' faces.
- You will be asked to rate them based on:
  - How trustworthy you think they look
  - How aggressive you think they look
  - How attractive you think they look
  - How friendly you think they look
  - And how wealthy you think they look
- **You will rate them on a scale from 1 to 10**
  - With 1 being **NOT VERY.**
  - And 10 being **VERY MUCH.**
- At the end of the experiment, the computer will pick one photo at random and one question at random.
  - **If your rating of that photo, for that question, is in line with the majority of other responses in the session, you will be paid £2.**
- Example:
  - Suppose the computer selects Picture 5, and selects the trustworthiness question. If you select a trustworthiness rating of 4 for Picture 5, and the **modal choice for that question (that is, the majority of other responses) for that photo is 4** you will receive **£2.**
  - If you selected a trustworthiness rating of 2, you will receive nothing.

C.1.5 Ex-post driver survey

UNIVERSITY OF EXETER

All questions are about passengers taken between 9am-5pm

1. How many passenger journeys do you normally complete between 9am and 5pm?

2. How many of those journeys start from a taxi rank?

3. What is the average fare for someone catching a taxi from a taxi rank?

4. What is the lowest fare you would accept for a journey starting from a taxi rank?

5. How many of those passengers taking a journey from a rank would leave a tip?

6. How much would they leave as a tip, on average?

7. How much do you earn per day, on average?

Consider a journey from **Manchester Piccadilly Station** to the **Coronation Street Tour**.

8. How much would you expect the fare for this journey to be?

9. Would you let a passenger bargain with you on the price of this journey before they entered the taxi?
   a. Yes
   b. No

10. **If yes**, what is the lowest fare you would accept for this journey?

11. Would you let a passenger bargain with you on the fare of this journey whilst you were driving the taxi?

    a. Yes
    b. No

12. If yes, what is the lowest fare you would accept for this journey?

13. Once you had completed this journey would you: **(Please circle one)**

    a. Return to Manchester Piccadilly

    b. Return to a different taxi rank. (Please state which one.)

    c. `Cruise' and look for a passenger to hail you down.

    d. Something different (please specify):

Consider a journey from **Manchester Piccadilly Station** to the **Stretford Mall**.

14. How much would you expect the fare for this journey to be?

15. Would you let a passenger bargain with you on the price of this journey before they entered the taxi?

    a. Yes
    b. No

16. If yes, what is the lowest fare you would accept for this journey?

17. Would you let a passenger bargain with you on the fare of this journey whilst you were driving the taxi?

    a. Yes
    b. N0

UNIVERSITY OF EXETER

18. If yes, what is the lowest fare you would accept for this journey?

19. Once you had completed this journey would you: **(Please circle one.)**

   a. Return to Manchester Piccadilly

   b. Return to a different taxi rank. Please state which one.

   c. `Cruise' and look for a passenger to hail you down?

   d. Something different (please specify):

<u>These questions are about you and your taxi</u>

1. **How old are you?**

2. **What is your gender?**

3. **What is your ethnicity?**

4. **Do you own your own taxi?**

5. **How old is the taxi you drive?**

6. **What is the make and model of your taxi?**

C.2 Statistical Appendix

C.2.1 Constructing the likelihood function

We assume the driver decides to stop based entirely on the taxi meter. As the meter increases in discrete amounts, the driver therefore makes a discrete choice: stop now, or wait until the next 'pulse' of the meter. This assumption seems reasonable, as each 'pulse' of the meter quantifies an exact distance driven. The driver must choose how many 'pulses' to give for free, x, bounded by the number of pulses until the journey is completed: $x \in \{0, 1, 2, ..., \bar{x}\}$, where $\bar{x}$ is the maximum number of pulses the driver can give for a given journey. When $x = \bar{x}$, the driver completes the journey.

Estimation begins from the observation that for any of the utility specifications outlined in Table 12, the driver's utility maximising choice of x, x^*, varies only with ϵ, the idiosyncratic error.

Fixing the model parameters, α, θ, a, b, c, d and e, we can determine the values of ϵ at which the driver's choice changes, ϵ_x. A driver will give x to the passenger over $x + 1$ until

$$u(x; \alpha, \theta, a, b, c, d, e, \epsilon_x) = u(x + 1; \alpha, \theta, a, b, c, d, e, \epsilon_x). \tag{C.1}$$

Taking the Cox *et. al* (2007) form as the example, $u(x) = [(s-x-g(x)\cdot p)^\alpha + \theta x^\alpha]\alpha^{-1}$, Equation C.1 can be rearranged as

$$\epsilon_x = \frac{(s - x - g(x) \cdot p)^\alpha - (s - x - g(x+1) \cdot p - 1)^\alpha}{(x+1)^\alpha - x^\alpha} - \theta,$$

where $\theta = \bar{\theta} \cdot (1 + a \cdot m_1 + b \cdot m_2 + c \cdot m_3 + d \cdot m_4 + e \cdot m_5)$, as defined in Section 5.3. Dividing through by σ gives,

$$\frac{\epsilon_x}{\sigma} = \frac{1}{\sigma} \left(\frac{(s - x - g(x) \cdot p)^\alpha - (s - x - g(x+1) \cdot p - 1)^\alpha}{(x+1)^\alpha - x^\alpha} - \theta \right). \tag{C.2}$$

When $\epsilon \in (\epsilon_{x-1}, \epsilon_x)$, then $x^* = x$; the probability of choosing x can therefore be determined from the cumulative distribution function of the error term. Where $f(z)$ is the density function, and $F(z)$ the cumulative distribution, the probability that the driver chooses $x^* = 0$ (i.e. stops at the amount the tester can afford) is the probability that $\epsilon \in (-\infty, \epsilon_0)$, or

$$\Pr[x^* = 0|\alpha, \theta, a, b, c, d, e, \sigma] = \int_{-\infty}^{\epsilon_0} f(z)dz = F(\epsilon_0). \tag{C.3}$$

The probability the driver chooses $x^* = q \in \{1, 2, ..., \bar{x} - 1\}$ is

$$\Pr[x^* = q|\alpha, \theta, a, b, c, d, e, \sigma] = \int_{\epsilon_{x-1}}^{\epsilon_x} f(z)dz = F(\epsilon_x) - F(\epsilon_{x-1}), \tag{C.4}$$

and the probability the driver completes the journey, $x^* = \bar{x}$, is

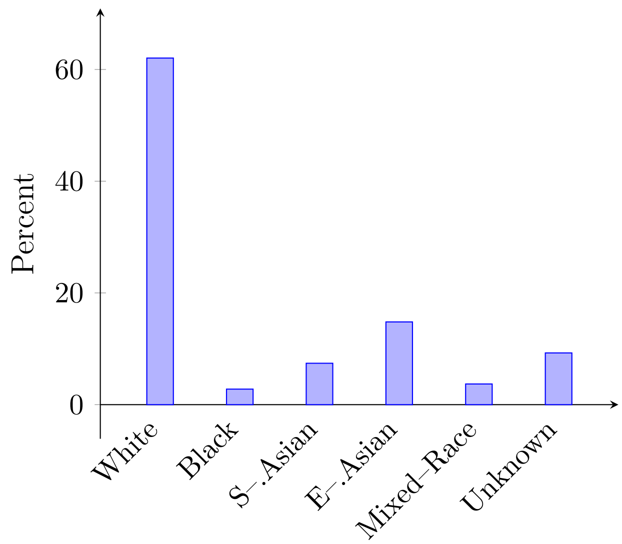
$$\Pr[x^* = \bar{x}|\alpha, \theta, a, b, c, d, e, \sigma] = \int_{\epsilon_{\bar{x}-1}}^{\infty} f(z)dz = 1 - F(\epsilon_{\bar{x}-1}). \tag{C.5}$$

The likelihood function, using 132 journeys from the *No Reputation* treatment, as specified in Section 5.3, is therefore

$$L(\alpha, \theta, a, b, c, d, e, \sigma) = \prod_{k=1}^{132} Pr[x_k = x; \alpha, \theta, a, b, c, d, e, \sigma]. \tag{C.6}$$

Taking logs gives the log–likelihood function, which can then be maximised with respect to the model parameters.

C.2.2 Raters' ethnic demographics



Note: 108 subjects took part in the rating task. The *Mixed–Race* category includes anyone who reported more than one ethnicity. The *Unknown* category includes those who did not report their ethnicity and those who reported an ambiguous ethnic affiliation.

Figure C.1: Distribution of the Raters' Self–Reported Ethnicity

C.2.3 Testers' appearance characteristics

Panel A: *Appearance Correlations, Pooled*					
	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.2839*	1.0000			
Friendly	-0.6507*	0.4358*	1.0000		
Trustworthy	-0.6641*	0.4319*	0.7835*	1.0000	
Wealthy	-0.3067*	0.4724*	0.3612*	0.3948*	1.0000
Note: 660 observations.					
Panel B: *Appearance Correlations, White Testers*					
	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.1501*	1.0000			
Friendly	-0.6051*	0.3766*	1.0000		
Trustworthy	-0.6189*	0.3383*	0.7746*	1.0000	
Wealthy	-0.2077*	0.5125*	0.2925*	0.3567*	1.0000
Note: 360 observations.					
Panel C: *Appearance Correlations, Black Testers*					
	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.4602*	1.0000			
Friendly	-0.6662*	0.5530*	1.0000		
Trustworthy	-0.6497*	0.6203*	0.7554*	1.0000	
Wealthy	-0.3814*	0.3674*	0.4039*	0.4190*	1.0000
Note: 210 observations.					
Panel D: *Appearance Correlations, S.Asian Testers*					
	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.2330*	1.0000			
Friendly	-0.5211*	0.2552*	1.0000		
Trustworthy	-0.5945*	0.2871*	0.6151*	1.0000	
Wealthy	-0.0586	0.2681*	0.0538	0.1408	1.0000
Note: 90 observations.					

Note: * indicates significance at the 5% level.

Table C.1: Tester Appearance Correlations

C.2.4 Reduced form ethnic interactions

Dep. Var:		Amount Given (£)				
Driver	*Tester*	(1)	(2)	(3)	(4)	(5)
White	Black	-0.565*	-0.609**	-0.605*	-0.605*	-0.662**
		(0.329)	(0.308)	(0.317)	(0.317)	(0.318)
White	Asian	0.282	0.236	0.226	0.226	0.407
		(0.424)	(0.394)	(0.409)	(0.409)	(0.409)
Asian	White	-0.276	-0.200	-0.247	-0.247	-0.239
		(0.210)	(0.211)	(0.216)	(0.216)	(0.214)
Asian	Black	-0.849***	-0.773***	-0.787***	-0.787***	-0.877***
		(0.243)	(0.231)	(0.251)	(0.251)	(0.247)
Asian	Asian	-0.487	-0.550**	-0.601**	-0.601**	-0.412
		(0.305)	(0.254)	(0.304)	(0.304)	(0.301)
Constant		0.775	1.248*	1.569**	1.569**	1.122
		(0.692)	(0.750)	(0.788)	(0.788)	(2.657)
Treatment Controls		✓	✓	✓	✓	✓
Driver Controls		✓	✓	✓	✓	✓
Tester Controls			✓	✓	✓	✓
Ride Controls				✓	✓	✓
Field Controls					✓	✓
City Controls					✓	✓
Appearance Controls						✓
Observations		255	254	254	254	254

Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. The estimates are obtained using observations from all treatments, but we exclude observations where the driver is black. The number of observations fall slightly as more controls are included due to missing entries. Appearance Controls include measures of the Testers' aggressiveness, attractiveness, friendliness, trustworthiness and wealthiness, as outlined in Section 4.4.1. Observations with a white driver and a white Tester are taken as the baseline.

Table C.2: The Effects of Ethnic Interactions on Giving

C.2.5 Robustness checks

#	*Alt. Hypothesis*	*Family*	*Outcome*	*Unadjusted*	*Adjusted*
Result 4					
1	H_A: White ≠ Black	*No–Reputation, Short*	Giving, £	0.001***	0.003***
2	H_A: White ≠ S.–Asian			0.37	0.37
3	H_A: S.–Asian ≠ Black			0.06*	0.12
4	H_A: White ≠ Black	*No–Reputation, Long*	Giving, £	0.396	1.00
5	H_A: White ≠ S.–Asian			0.88	0.88
6	H_A: S.–Asian ≠ Black			0.47	0.94
7	H_A: White ≠ Black	*No–Reputation, pooled*	Giving, %	0.005***	0.015**
8	H_A: White ≠ S.–Asian			0.311	0.311
9	H_A: S.–Asian ≠ Black			0.025**	0.05**
10	H_A: White ≠ Black	*Reputation, Short*	Giving, £	0.0003***	0.0009***
11	H_A: White ≠ S.–Asian			0.13	0.26
12	H_A: S.–Asian ≠ Black			0.622	0.622
13	H_A: White ≠ Black	*Reputation, Long*	Giving, £	0.566	1.00
14	H_A: White ≠ S.–Asian			0.46	0.46
15	H_A: S.–Asian ≠ Black			0.45	0.90
16	H_A: White ≠ Black	*Reputation, pooled*	Giving, %	0.0005***	0.0015***
17	H_A: White ≠ S.–Asian			0.003***	0.006***
18	H_A: S.–Asian ≠ Black			0.90	0.90
Result 5					
19	H_A: White ≠ Black	*No–Reputation, pooled*	Journey Completion	0.045**	0.135
20	H_A: White ≠ S.–Asian			0.793	0.793
21	H_A: S.–Asian ≠ Black			0.088*	0.176
Result 6					
22	H_A: *No Rep.* ≠ *Rep.*, White	*No–Rep. vs Rep., Short*	Giving, £	0.055*	0.165
23	H_A: *No Rep.* ≠ *Rep.*, Black			0.593	0.593
24	H_A: *No Rep.* ≠ *Rep.*, S.–Asian			0.278	0.556
25	H_A: *No Rep.* ≠ *Rep.*, White	*No–Rep. vs Rep., Long*	Giving, £	0.61	1.00
26	H_A: *No Rep.* ≠ *Rep.*, Black			0.624	0.624
27	H_A: *No Rep.* ≠ *Rep.*, S.–Asian			0.506	1.00
28	H_A: *No Rep.* ≠ *Rep.*, White	*No–Rep. vs Rep., pooled*	Giving, %	0.053*	0.159
29	H_A: *No Rep.* ≠ *Rep.*, Black			0.397	0.397
30	H_A: *No Rep.* ≠ *Rep.*, S.–Asian			0.15	0.3

Note: ***, ** and * represent significance at the 1%, 5% and 10% levels. *Adjusted* p-values are adjusted using the Holm–Bonferroni procedure. All tests are two sided.

Table C.3: Adjusted p-values – Non–Parametric Testing

Table	#	*Result*	*Model*	*Explanatory Variable of Interest*			
				Black	*South–Asian*	*Male*	m
	1		(1)	0.002***	0.615		2
	2		(2)	0.003***	0.279	0.106	3
Table 4.10	3	Result 4	(3)	0.000***	0.129	0.06*	3
	4		(4)	0.005***	0.287	0.126	3
	5		(5)	0.001***	0.981	0.462	8
	6		(1)	0.022**	0.197		2
	7		(2)	0.048**	0.153	0.246	3
Table 4.11	8	Result 5	(3)	0.049**	0.312	0.286	3
	9		(4)	0.057*	0.242	0.188	3
	10		(5)	0.056*	1.00	1.00	8

Note: ***, ** and * represent significance at the 1%, 5% and 10% levels. *Adjusted* p-values are adjusted using the Holm–Bonferroni procedure. All tests are two sided. Column m outlines how many comparisons were made within the family of hypotheses.

Table C.4: Adjusted p–values – Parametric Testing

Appendix D

Retrospective Power Analysis

As we fail to find statistically significant differences stemming from treatments in Chapter 2, Chapter 3 and Chapter 4, we conduct a retrospective power analysis. This is done in order to examine the probability that the observed null results are a consequence of Type II error: the probability that we failed to reject the null hypothesis even though it is false (a 'false negative'). Although there is a large and heated debate surrounding the usefulness and interpretation of post-hoc power analysis, there is no doubt that it is useful for informing future research of required sample sizes (Hoenig & Heisey, 2001).

For each chapter, first, we calculate the statistical power of the experiment assuming the observed effect size in our sample is equal to the effect size in the population. Second, we calculate the power of the experiment if the observed effect size had been *Medium* and *Large*, as defined by Cohen (1992). This is done using the number of observations and standard deviations observed in each of the respective samples. This will shed light on the extent to which our experimental planning regarding sample sizes is justified or not.

Table D.1 outlines the calculated statistical power for each experiment and for each effect size. Note, that the power to detect the observed treatment effect is 20% or smaller in each case. This is to be expected, given the high calculated p–values, and the one to one relationship between statistical power and calculated p–values (Hoenig & Heisey, 2001).[1] Further, for each chapter, Cohen (1992) would class the observed treatment differences as *Small*. Had we observed *Medium* or *Large* effects in each chapter conditional on the sample standard deviations being representative, as Table

[1]It is important to note that, in an experiment where a treatment comparison has a calculated p–value of $p = 0.05$, the experiment is powered at only 50% (Hoenig & Heisey, 2001).

	Observed Effect	*Medium Effect**	*Large Effect***
Chapter 2^{+}	20%$^{++}$	99%	99%
Chapter 3$^{\Lambda}$	20%$^{\Lambda\Lambda}$	75%	93%
Chapter 4^{o}	14%^{oo}	98%	99%

**Medium* effect size, d=0.5. ***Large* effect size, $d = 0.8$.
$^{+}$Power calculated for the comparison of *Bonus* payments between the *In–group* and *Out–group* treatments. $^{++}$Observed effect size, $d = 0.13$.
$^{\Lambda}$Power calculated for the comparison of dictator giving between the *Anonymous* and *English* treatments. $^{\Lambda\Lambda}$Observed effect size, $d = 0.27$.
^{o}Power calculated for the comparison of giving as a percentage of the expected fare between the *No Reputation* and *Reputation* treatments. ^{oo}Observed effect size, $d = 0.1$.

Table D.1: Retrospective Power Analysis by Chapter

D.1 outlines, the experiments would have been powered at conventionally accepted levels.