Do you mind me paying less?
Measuring Other–Regarding Preferences in the Market for Taxis*

Brit Grosskopf[1] and Graeme Pearce[2]

[1] *University of Exeter*
[2] *University of Innsbruck*

November 8, 2017

Abstract

We present a natural field experiment designed to measure other–regarding preferences in the market for taxis. 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–subject design we vary the length of the journey and whether or not a business card is elicited. We find that the majority of drivers give at least part of the journey for free, that giving is proportional to the length of the journey and that over 25% of drivers complete the journey. 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 estimate the parameters of a number of utility functions. The data and the structural analysis lend support to the quantitative predictions of experiments that measure other–regarding preferences, and shed further light on how discrimination can manifest itself within our preferences.

*We are grateful to seminar participants at Texas A&M University, the University of Exeter brown bag, Royal Holloway University of London, Humboldt University, and attendees at the ESA North American meetings in Dallas and the 2016 Royal Economic Society meetings for helpful comments and suggestions. We also thank Loukas Balafoutas and Henry Schneider for comments. We thank the University of Exeter Business School for funding this research.

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 market 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 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 received similar criticisms to experiments studying social preferences (Zizzo, 2010, 2012; Guala & Filippin, 2015).

The purpose of this paper is to examine the prevalence, and extent, of other–regarding preferences in a highly competitive market place. We also investigate the significance of ethnic identity in determining these preferences. 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.[2]

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

[2] The taxi markets we study satisfy all the requirements of a market place, as discussed by Al-Ubaydli & List (2016).

In a 2×2 between–subjects design we systematically vary the length of the taxi journeys using *Short* and *Long* distance treatments, where testers take journeys of approximately 1.7 miles and 4.4 miles. Since drivers assigned to the *Long* distance treatment 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. 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, and the only real possibility of meeting a driver in a future interaction is by obtaining his/her contact details so that he/she can be actively selected. In the *Baseline* treatments, testers reinforce the one–shot nature of the interaction by stating to the driver that they, "don't take taxis very often." However, as shown by List (2006), field experiments designed to detect social preferences need to be particularly careful about the possibility of reputational concerns. To address this, and to reinforce the one–shot nature of the *Baseline* treatment, we conduct a *Business Card* treatment to examine if drivers are willing to give out their contact details for potential future interactions.

We find that 70% of drivers in the *Baseline* 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, with drivers giving around 10% of the expected fare in both *Short* and *Long* distance treatments. In addition, drivers do not seem concerned about repeated business with customers, with a minority (45%) producing a business card when asked to do so. Drivers who fail to give a business card behave identically to those assigned to the *Baseline* treatment. This demonstrates the inherent one–shot nature of the interactions in the market we study, and we feel confident that the drivers' behaviour from the *Baseline* treatments is not influenced by reputational concerns. However, drivers assigned to the *Business Card* treatment who do provide a business card are found to give significantly more than those who do not, but only for *Short* distance journeys.

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 behavioural 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.

This study makes a number of contributions. First, we contribute to the debate on

the generalisability of laboratory experiments by providing the first evidence that other–regarding preferences can appear in a natural competitive market setting with a similar prominence to that observed in the laboratory.[3] 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 demonstrate that the drivers' behaviour in the *Baseline* treatments cannot be attributed to reputational concerns. 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 2 reviews the relevant literature, Section 3 discusses the taxi markets we study and Section 4 outlines the experimental design in detail. Section 5 outlines reduced form estimation results, and estimates from a structural model. Section 6 examines the robustness of our results. Section 7 discusses alternative interpretations of the results and Section 8 concludes.

2 Literature

2.1 Other–regarding preferences

As highlighted in the reviews of Camerer & Fehr (2004) and Cooper & Kagel (2009), pro–social preferences are well established to exist in the laboratory, with experiments typically reporting high levels of other–regarding behaviours. However, the extent of other–regard has been found to be highly dependent on the experimental procedure (Haley & Fessler, 2005; Hoffman *et al.*, 1996). Further, individuals found to have such preferences are often observed to behave selfishly under different institutions, with competitive settings 'crowding out' other–regarding behaviour. For example, individuals often reject 'unfair' offers in ultimatum games, and give positive amounts in dictator games suggesting that subjects are inequality averse (Engel, 2011; Fehr & Schmidt, 1999). Yet, many experimental markets converge on the competitive equilibrium.[4] Whilst some suggest this result is indicative that individuals do not have other–regarding preferences, the models of Fehr & Schmidt (1999) and Bolton & Ockenfels (2000) predict this outcome. Institution–dependent behaviour could be explained by individuals being unable to enforce an equitable outcome within a market setting (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. However, Dufwenberg *et al.* also show that if individuals have these preferences, market allocations are not necessarily efficient.[5]

[3] There is a rich literature examining reciprocity using both laboratory and natural field experiments (Falk, 2007; Kube *et al.*, 2012; Gneezy & List, 2006). Although a social preference, reciprocity is not an other–regarding preference, with a number of studies showing that reciprocity is not driven by a concern for others (see Falk & Fischbacher (2006) for a comprehensive discussion of the determinants of reciprocal actions).

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

[5] Schmidt (2011) provides an excellent review of this literature, and Al-Ubaydli & List (2016) provides a discussion of the efficiency implications of welfare externalities.

In order to examine these findings outside the laboratory, field experimenters have had to overcome two challenges. First, in order to address the generalisability and interpretability problems associated with the study of pro–social behaviours in the lab (Levitt & List, 2007; Zizzo, 2012), field experimenters had to seek a variety of field contexts that mirror the social dilemmas faced by laboratory subjects (Stoop *et al.*, 2012; Stoop, 2014). Second, due to the possibility of repeated interaction that is inherent in the field, field experimenters have had to be particularly careful about the possibility of reputational and monitoring concerns explaining behaviour that is otherwise consistent with social preferences.

This was first raised by List (2006), who 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 reciprocal behaviour, 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). However, there are lab and field studies that 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 transactions where customers can evaluate how favourable the service is. In other transactions, reputation has no effect. The authors highlight how the predicted effect of reputation on behaviour will depend 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, the latter is likely to surpass the former. 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 argument that reputation and social preferences are substitutes.

In some of the first evidence of other–regarding preferences in a social setting in the field, DellaVigna *et al.* (2012) use a novel natural field experiment nested within a charitable door–to–door fund raiser. The authors are able to disentangle altruism from social pressure effects by giving 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.[6] However, as far as we are aware, there exists no field study in which other–regarding preferences have been identified in a competitive market setting.

2.2 Discrimination

As Bertrand & Duflo (2016) highlight, there is considerable evidence that discrimination against minority groups is pervasive. Although economic research has predominantly focused on discrimination in labour markets, with specific attention given to parsing discrimination into statistical and taste–based explanations, experimenters have also examined consumer goods markets (Ayres *et al.*, 2015; Doleac & Stein, 2013) and helping behaviours (Gneezy *et al.*, 2012). Economists typically employ either audit studies or correspondence

[6] This is in line with previous studies that find that social pressure has a significant effect on pro–social behaviour (Gerber *et al.*, 2008; Mas & Moretti, 2009).

studies in order to detect discriminatory behaviour.[7]

Those studies which are most related to ours, audit studies, utilise actors to take part in standardised interactions such as job interviews or negotiations (Ayres & Siegelman, 1995; Castillo *et al.*, 2013).[8] These studies have typically 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 field 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.

In a correspondence study, the experimenter fabricates CVs or letters whilst varying either the ethnicity, nationality or gender of the applicant or sender 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 studies, and findings, have been reported in Australia (Booth *et al.*, 2012), Canada (Oreopoulos, 2011) and Israel (Ruffle & Shtudiner, 2014). Such studies come close to identifying a 'taste' for discrimination, although statistical discrimination can often not be ruled out.

[7] 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.

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

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. 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 themselves. Our study distinguishes itself from Mujcic & Frijters (2013) in a number of important and economically significant ways. First, our subjects are not monitored and are not observed by third parties or by an employer. Second, we consider behaviour in a competitive context where the subjects' objective function is typically assumed to be their profits. This contrasts with Mujcic & Frijters (2013), who study a social context where the decision maker is not concerned with their own profits. Finally, our study elicits other–regarding preferences more generally, considering discrimination in a situation where pro–social behaviour is *costly* to the person exhibiting it.

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.

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.[9] 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

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

		Greater Manchester		
	Birmingham	Manchester	Trafford	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

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 fares based on a journey made by a single passenger with no luggage, between 9am and 5pm.

Table 1: Taxi Fares by Local Authority

authority. Those relevant for this study are detailed in Table 1.

Important to our study is the fact that the metered fare is the maximum fare the driver is able to charge the passenger. 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 driver owned: drivers keep all the fare, any tips (which are typically around 10%), and incur all the costs associated with a journey.[10] 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 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 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. 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.

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

			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 2: Taxis, Taxi Ranks and Weekly Passenger Numbers

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 A).[11] We hired 22 testers in total. This compares favourably to previous studies of taxi markets that have typically employed just a handful of testers.[12]

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.[13] 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

[11] In line with the ethics guidelines at the University of Exeter, we invited everyone who applied for the job to an interview and made job offers to everyone who attended the interview.

[12] For example, Balafoutas *et al.* (2013) employed five testers.

[13] We discussed the *Short* distance / *Baseline* treatment, which is described in Section 4.2.

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.[14] 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 (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 characteristic 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 3.[15,16]

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,

[14] 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.

[15] Table 19, in Appendix B, presents the correlations between the testers' perceived facial appearance characteristics.

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

		Tester Ethnicity		
	All testers	White	Black	S.–Asian
Age	27.57	29.45	26.14	24
	(8.25)	(10.18)	(5.64)	(4.58)
Gender (1 if male)	0.68	0.58	0.86	0.67
	(0.48)	(0.51)	(0.38)	(0.58)
Aggressiveness	4.04	3.99	4.63	2.82
	(1.52)	(1.5)	(1.61)	(0.71)
Attractiveness	4.77	4.86	4.42	5.2
	(1.41)	(1.62)	(1.3)	(0.69)
Friendliness	5.91	5.85	5.5	7.13
	(1.68)	(1.69)	(1.79)	(1.25)
Trustworthiness	5.68	5.69	5.18	6.81
	(1.49)	(1.5)	(1.54)	(1.01)
Wealthiness$^{\diamond}$	5.27	5.45	4.56	6.21
	(1.14)	(1.27)	(0.68)	(0.19)
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 19 in Appendix B. The raters' ethnicities are presented in Figure 4 in Appendix B.

$^{\diamond}$ 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 3: Tester Characteristics

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 Order Test). We control for these tester specific variations in our parametric analysis in Section 5.

We focus on facial appearance due to the way that the driver and tester interact whilst in the taxi. As outlined in Section 4.2, 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.

4.2 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 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

		Short Distance	*Long Distance*
Baseline	Entry Script	*"I don't take taxis very often."*	
	Endowment	£4	£8
	Expected Fare	£5	£10
Business Card	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: Experimental Design Summary

£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. Testers took part in multiple treatments across different days.

Upon entering the taxi, the tester first stated their destination, and then spoke a simple entry statement.[17] In the *Baseline* treatment they stated, "I don't take taxis very often", and in the *Business Card* 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 gave the driver the opportunity to give out a business card before the journey began. The proportion of cards given out should provide some indication of the drivers' concern for repeated business, with a higher proportion signalling a higher concern. 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 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 summarises the experimental design.

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

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

	Driver Ethnicity				
	All Drivers	White	Black	South–Asian	Other
Age	44.34	50.06	40.36	42.6	41.33
	(10.67)	(10.56)	(9.36)	(10.03)	(11.45)
Gender (1 if male)	0.99	0.97	1	1	1
	(0.12)	(0.17)	(0)	(0.1)	(0)
Journeys	283	71	11	191	10

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 the driver outside the 3 main ethnic groups that are specified.

Table 5: Driver Characteristics

Field Characteristics	
Traffic (1 if Not Busy, 10 if Very Busy)	4.44
	(2.26)
Weather (1 if raining, 0 otherwise)	0.11
	(0.32)
Ride Characteristics	
Conversation (1 if driver attempted a conversation)	0.28
	(0.45)
Cashpoint (1 if driver offered a cashpoint)	0.04
	(0.2)
Business card, Business Card treatment only (1 if given)	0.22
	(0.42)
Receipt (1 if given)	0.89
	(0.31)

Note: Mean averages of each variable. Standard deviations in parentheses.

Table 6: Taxi Journeys Characteristics

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/ATM (1 if yes) and (in the *Business Card* 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.[18] Table 5 presents driver characteristics, as reported by the testers.[19] Table 6 reports mean values

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

[19]Although Manchester Council do not collect driver ethnic demographics, Birmingham City Council provided the following information regarding the distribution of driver ethnicities (obtained from a freedom of Information request, number FOI 15327): 82.6% South–Asian, 9.6% white, 3.8% black and 3.9% other.

and standard deviations of all the other variables recorded by the testers.

At this point, 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 make. 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.

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.

5.1 Journey calibration checks

Some initial calibration checks are conducted in order to examine if our expected fare calculations are accurate. Table 7 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 of a completed journey to its expected fare, the accuracy of our expected fare calculations can be examined. This will also shed light on how the drivers perceived the testers, i.e. as locals or non-locals (Balafoutas *et al.*, 2013). Minor discrepancies 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 ($p = 0.304$) or *Long* distance treatment ($p = 0.539$) are reported. The amount given as a percentage of the expected fare is not significantly different to the planned 20% in both the *Short* ($p = 0.88$, Sign Test) and *Long* ($p = 1$, Sign Test) distance treatments. We conclude that our testers were perceived as locals by the drivers and that our journey planning is accurate.

5.2 Other–regard

Table 8 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 1 displays the distribution of giving across treatments.

Other includes all drivers who declared other ethnicities (e.g. mixed), and those who did not disclose their ethnicity. The ethnic distribution of our sample is representative of the population distribution.

	Short Distance	*Long Distance*
Recorded Fare (£)	5.44	10.43
	(1.3)	(1.47)
Expected Fare (£)	5.65	10.22
	(0.22)	(0.93)
Amount Given as a % of the expected fare	0.26	0.24
	(0.23)	(0.15)
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 7: Fares, Expected Fares and Average Giving Conditional on the Driver Completing the Journey

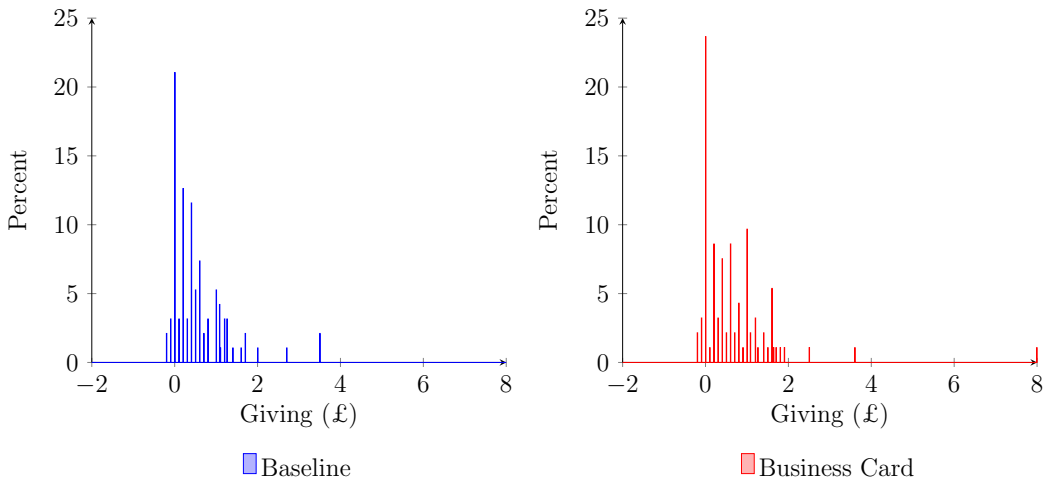
	Baseline		*Business Card*	
	Short	*Long*	*Short*	*Long*
Amount Given (£)	£0.6	£1.14	£0.72	£1.08
	(0.73)	(1.43)	(1.07)	(1.23)
Amount Given as a % of the Expected Fare	0.11	0.11	0.13	0.1
	(0.13)	(0.15)	(0.19)	(0.12)
Proportion of Completed Journeys	0.27	0.27	0.31	0.34
Number of Journeys	95	48	93	47

Note: Standard deviations in parentheses.

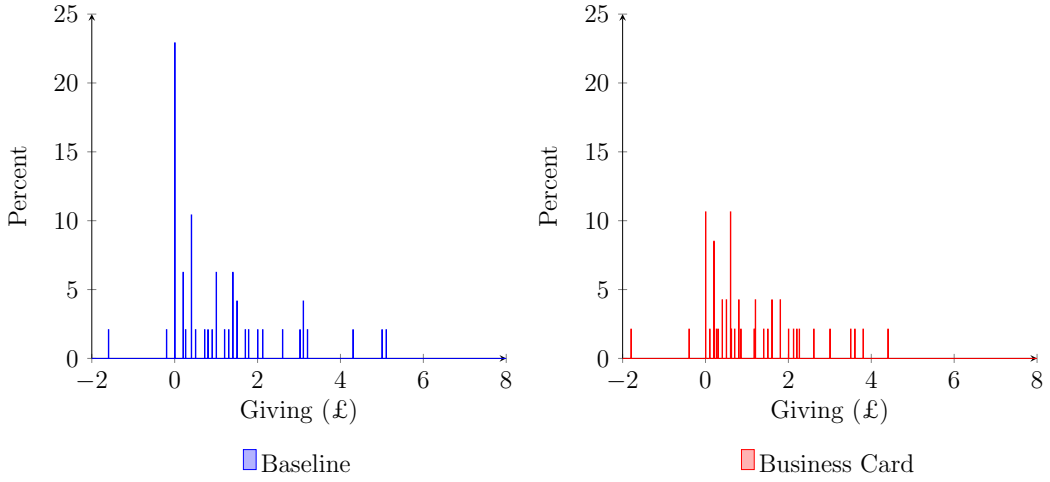
Table 8: Average Driver Giving, by Treatment

Table 9 reports a number of random effects Tobit regressions. In models (1), (2), (3) and (4) giving in pounds by driver i to tester j is the dependent variable. In models (5), (6), (7) and (8), 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 *Business Card* treatment are included along with their interaction; the *Short* distance *Baseline* treatment is taken as the control. In models (4) and (8) we include an additional dummy variable, *B.C. Received* along with its interaction with the *Long* distance treatment variable, to examine the correlation between driver giving behaviour and the decision to provide a business card (1 if provided, 0 otherwise). Thus, the coefficient on *B.C. Received* captures the correlation between giving and a business card being provided in the *Short* distance *Business Card* treatment.

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 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



(a) Short Distance Journeys



(b) Long Distance Journeys

Figure 1: Distribution of Giving, by Treatment

for the weather conditions (1 if raining). The set of 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. The set of Ride Controls includes dummies controlling for whether the driver offered to take the passenger to a cashpoint/ATM (1 if offered) and if he tried to engage in a conversation (1 if yes). All regressions include tester fixed effects.

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

Support. Considering journeys from the *Baseline* 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 9, with a positive and significant constant in all regression models except (6) ($p < 0.05$, in all cases).

	Random Effects Tobit Regressions							
Dep. Variable:	Amount Given (£)				Amount Given as a % of the Exp. Fare			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Long	0.676***	0.729***	0.734***	0.754***	0.049	0.07	0.07	0.075
	(0.254)	(0.278)	(0.272)	(0.275)	(0.068)	(0.075)	(0.074)	(0.074)
B.Card	0.072	0.042	0.048	-0.145	0.038	0.03	0.033	-0.047
	(0.157)	(0.159)	(0.156)	(0.189)	(0.042)	(0.042)	(0.042)	(0.05)
B.Card × *Long*	0.034	0.062	0.034	0.156	0.011	0.026	0.023	0.083
	(0.295)	(0.301)	(0.295)	(0.334)	(0.078)	(0.08)	(0.08)	(0.089)
B.C. Recieved				0.414*				0.178***
				(0.232)				(0.062)
B.C. Recieved × *Long*				-0.293				-0.159
				(0.418)				(0.111)
Constant	1.243***	1.495***	1.388***	1.289***	0.293***	0.335	0.31***	0.285**
	(0.389)	(0.389)	(0.06)	(0.44)	(0.1)	(0.389)	(0.018)	(0.114)
Observations	283	282	281	280	283	282	281	280
City Controls	✓	✓	✓	✓	✓	✓	✓	✓
Field Controls		✓	✓	✓		✓	✓	✓
Ride Controls			✓	✓			✓	✓

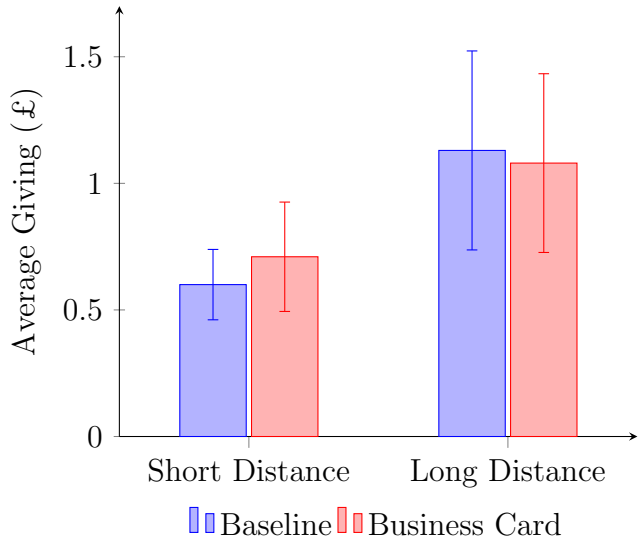
Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. The number of observations fall slightly as more controls are included due to missing entries. Models (1), (2), (3) and (4) are left censored at 0, and right censored at the difference between the expected fare had the driver completed the journey, and the amount paid by the tester. Models (5), (6), (7) and (8) are left censored at 0, and right censored at 1. All regressions include tester fixed effects.
Variables: Long, 1 if a *Long* distance journey, 0 otherwise; *B. Card*, 1 if assigned to the *Business Card* treatment, 0 otherwise. *Received*, 1 if a business card was provided, 0 otherwise.

Table 9: Treatment Effects

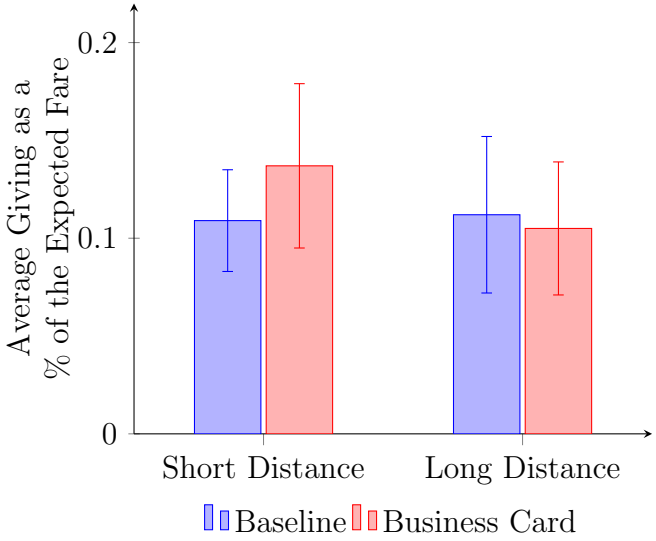
Similar findings are observed in the *Business Card* treatment, with over 75% of drivers giving at least part of the journey for free, and 32% of all journeys being completed in full.

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

Support. Examining journeys from the *Baseline* treatment, average driver giving is significantly different in *Short* distance journeys in comparison to *Long* distance journeys ($p = 0.079$, Robust Rank Order Test). This is shown graphically in Figure 2a. The distribution of giving is also found to vary by the distance of the journey ($p = 0.059$, Kruskal–Wallis Test). Table 9, regressions (1), (2), (3) and (4) support these conclusions, reporting significant and positive coefficient estimates on the *Long* distance dummy ($p < 0.01$), whilst the coefficient on the *Business Card* 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 ($p = 0.88$, Robust Rank Order Test) (see Figure 2b). Further, the distance of the journey has no significant effect on its distribution ($p = 0.86$, Kruskal–Wallis Test). Estimates from Table 9 models (5), (6), (7) and (8) support this conclusion;



(a) Average Giving (£)



(b) Average Giving as a Percentage of the Expected Fare
Note: Vertical bars represent 95% confidence intervals.

Figure 2: Average Giving

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. *Asking for a business card has no effect on drivers' behaviour.*

Support. Comparing average giving between *Business Card* and *Baseline* treatments, no significant differences are reported in either *Short* or *Long* distance journeys ($p = 0.34$ and $p = 0.67$, Robust Rank Order Tests). Similarly, asking for a business card has no significant impact on the distribution of giving in either *Short* or *Long* distance treatments ($p = 0.67$ and $p = 0.44$, Kruskal–Wallis Test). The same is true for giving as a percentage of the expected fare, with no significant differences found between *Business Card* and *Baseline* 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 9 support these results, with the coefficient on the *Business Card* dummy found to be not significant at conventional levels across regressions ($p > 0.1$ in all cases). The estimates from Table 9 also outline how drivers that do select into giving a business card do not give significantly more of the journey for free in absolute terms than those in the *Baseline*, with the coefficient on both *B.C. Received* and on *B.C. Received* interacted with the *Long* distance journey dummy not found to be significantly different to zero in model (4) ($p > 0.1$). Interestingly, drivers that do provide a business card do give significantly more than those in the *Baseline* as a percentage of the expected fare in the *Short* distance journeys ($p < 0.01$, model 8).

Result 3 outlines how drivers' behaviour is, on average, unaffected by the *Business Card* treatment. This is most likely because the majority of drivers choose not to provide a business card, with only 45% deciding to provide one when asked (see Table 6). However, we acknowledge that this does not mean that repeated interaction effects are non–existent in this market place. Rather, we interpret Result 3 as being supportive of the idea that drivers who ply for hire at taxi ranks treat the interactions as one shot, and thus that our *Baseline* interactions are not confounded with reputational concerns. Further, the positive correlation between providing a business card and giving is suggestive that repeated interaction effects could play a role in fostering other–regard. Interestingly, the drivers who opt into a potential repeated interaction do so at the smallest possible cost, i.e. they only give more in the *Short* distance journeys, underscoring the strategic nature of reputation. After giving a business card, it is ambiguous that the driver should opt for increasing his reputation through giving. Drivers may 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, the drivers may 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 a '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 10 and Figure 3 displays the proportion of completed journeys by tester ethnicity graphically. To determine the effect of the testers' ethnicity on the drivers' behaviour, Table 11 outlines the results from a number of random effects Tobit regressions. In each case, giving as a percentage of the expected fare by driver

		Treatment			
		Baseline		*Business Card*	
Ethnicity		*Short*	*Long*	*Short*	*Long*
	Amount Given (£)	£0.64	£1.23	£0.87	£1.22
		(0.62)	(1.44)	(0.74)	(1.31)
White	Amount Given, % Exp. Fare	11%	13%	16%	12%
		(0.11)	(0.16)	(0.13)	(0.13)
	Journeys	60	26	49	29
	Amount Given (£)	£0.28	£0.79	£0.57	£1.05
		(0.46)	(0.99)	(1.57)	(1.31)
Black	Amount Given, % Exp. Fare	5%	8%	10%	10%
		(0.08)	(0.09)	(0.28)	(0.12)
	Journeys	26	11	30	11
	Amount Given (£)	£1.23	£1.28	£0.52	£0.54
		(1.4)	(1.8)	(0.65)	(0.53)
South–Asian	Amount Given, % Exp. Fare	23%	12%	9%	5%
		(0.26)	(0.17)	(0.12)	(0.05)
	Journeys	9	11	14	7

Note: Standard deviations given in parentheses.

Table 10: Summary Statistics, by Tester Ethnicity.

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 *Business Card*) and their interaction, along with a dummy controlling for a business card being provided by the driver (1 if provided, 0 otherwise), which is interacted with the treatment dummies. 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 sets described for Table 9. 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 11.

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

Support. Pairwise comparisons of average giving by drivers to white, black and South–Asian testers in the *Baseline* treatment reveals no significant differences between white and South–Asian testers in the *Short* or *Long* distance treatments ($p = 0.41$ and $p = 0.88$, 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 $p = 0.37$,

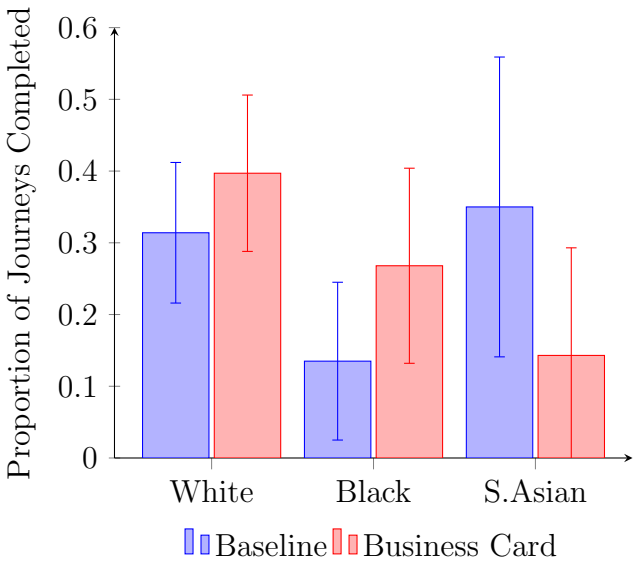
Random Effects Tobit Regressions					
Dep. Variable:	Amount Given as a % of the Expected Fare				
	(1)	(2)	(3)	(4)	(5)
Black	-0.15***	-0.122***	-0.123***	-0.112**	-0.129***
	(0.046)	(0.046)	(0.045)	(0.045)	(0.049)
South–Asian	-0.052	-0.061	-0.067	-0.06	-0.008
	(0.06)	(0.056)	(0.055)	(0.055)	(0.066)
Male		-0.117***	-0.115***	-0.108**	-0.15**
		(0.041)	(0.042)	(0.043)	(0.063)
Aggressiveness					0.005
					(0.069)
Attractiveness					0.026
					(0.03)
Friendliness					0.018
					(0.05)
Trustworthiness					-0.032
					(0.035)
Wealthiness					-0.042
					(0.072)
Constant	0.15	0.278	0.258	0.345*	0.542
	(0.177)	(0.19)	(0.189)	(0.196)	(0.746)
Observations	274	274	273	273	273
Treatment Controls	✓	✓	✓	✓	✓
Driver Controls	✓	✓	✓	✓	✓
Tester Controls		✓	✓	✓	✓
Ride Controls			✓	✓	✓
Field Controls				✓	✓
City Controls				✓	✓

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 1. Treatment Controls include dummies for each of the treatments (1 if *Long*, and 1 if *Business Card*) and their interaction, along with a dummy controlling for a business card being provided by the driver (1 if provided, 0 otherwise), which is interacted with the treatment dummies.
Variables: Black, 1 if black, 0 otherwise; *South–Asian*, 1 if South–Asian, 0 otherwise; *Male*, 1 if male, 0 otherwise; *Aggressiveness*, *Attractiveness*, *Friendliness*, *Trustworthiness* and *Wealthiness* are variables that take values on a 10 point Likert Scale.

Table 11: The Determinants of Driver Giving

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 ($p = 0.07$ and $p = 0.47$, Robust Rank Order Tests).

Considering 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 ($p = 0.002$, $p = 0.045$, Robust Rank Order Tests), but no differences are found between white and South–Asian testers ($p = 0.56$, Robust Rank Order Test). The estimates in Table 11



Note: Vertical bars represent 95% confidence intervals.

Figure 3: Proportion of Journeys Completed, by Tester Ethnicity

further support the non–parametric results: across all regressions, the coefficient on the black dummy is negative, highly significant ($p < 0.05$ in all cases, Wald Tests), and robust to changes in the model specification.

The differential treatment of testers by ethnicity remains in the *Business Card* treatment, with white testers receiving more than black testers in the *Short* distance treatment ($p < 0.001$, 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 $p = 0.02$, Robust Rank Order Tests), but no difference between black and South–Asian testers ($p = 0.622$, Robust Rank Order Test).

The proportion of completed journeys, by tester ethnicity, is now considered. Table 12 reports the estimated coefficients and marginal effects from 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 11.

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 *Baseline* treatment ($p = 0.045$ and $p = 0.088$, Fisher's Exact Test). No significant differences are reported between white and South–Asian testers

Random Effects Probit Regressions					
Dep. Variable:	Journey Completed				
	(1)	(2)	(3)	(4)	(5)
Black	-0.494**	-0.509**	-0.544**	-0.512**	-0.65***
	(0.198)	(0.217)	(0.221)	(0.223)	(0.246)
South–Asian	-0.313	-0.419	-0.45*	-0.438	-0.124
	(0.242)	(0.267)	(0.268)	(0.271)	(0.326)
Male		-0.335*	-0.335*	-0.326	-0.344
		(0.19)	(0.193)	(0.201)	(0.31)
Aggressiveness					-0.222
					(0.344)
Attractiveness					0.183
					(0.149)
Friendliness					-0.239
					(0.244)
Trustworthiness					0.022
					(0.342)
Wealthiness					-0.332*
					(0.172)
Constant	-0.913	-0.152	-0.074	-0.293	-03.658
	(0.76)	(0.862)	(0.877)	(0.919)	(3.621)
Observations	274	274	273	273	273
Marginal Effects					
Black	-0.16***	-0.161**	-0.169***	-0.159**	-0.188***
	(0.06)	(0.065)	(0.064)	(0.065)	(0.064)
South–Asian	-0.107	-0.136*	-0.143*	-0.139*	-0.041
	(0.078)	(0.081)	(0.079)	(0.079)	(0.106)
Treatment Controls	✓	✓	✓	✓	✓
Driver Controls	✓	✓	✓	✓	✓
Tester Controls		✓	✓	✓	✓
Ride Controls			✓	✓	✓
Field Controls				✓	✓
City Controls				✓	✓

Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. Marginal effects evaluated where the passengers' ethnicity is white, with all other variables evaluated at the mean. *Treatment Controls* includes a dummy for *Long* distance journeys, a dummy for the *Business Card* treatment and a dummy for if a business card was received, along with interactions.
Variables: Black, 1 if black, 0 otherwise; *South–Asian*, 1 if South–Asian, 0 otherwise; *Male*, 1 if male, 0 otherwise; *Aggressiveness*, *Attractiveness*, *Friendliness*, *Trustworthiness* and *Wealthiness* are variables that take values on a 10 point Likert Scale.

Table 12: Determinants of Journey Completion

($p = 0.793$, Fisher's Exact Test). The results from the random effects Probit regressions in Table 12 outline how the estimated coefficient on the black dummy is negative and signifi-

cant ($p = 0.05$). This estimate is robust to specification changes, and becomes increasingly significant as more controls are included. The estimated marginal effect size is robust across models, and is estimated to be highly significant ($p < 0.05$, in all cases). Similar to the coefficient estimates in Table 11, none of the appearance characteristics are significant.

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 4.1. 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: *Asking for a business card increases driver giving when the tester is white, has no effect when the tester is black and reduces giving when the tester is South–Asian.*

Support. White testers are given significantly greater amounts as a percentage of the expected fare in the *Business Card* treatment compared to the *Baseline* treatment for *Short* distance journeys ($p = 0.048$, Robust Rank Order Test, One sided). 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 asking for a business card ($p > 0.1$ in all cases, Robust Rank Order Tests). South–Asian testers see no effect from asking for a business card on giving in both the Short ($p = 0.27$, Robust Rank Order Test) and *Long* distance treatments ($p = 0.5$, Robust Rank Order Test). A negative effect is reported in giving as a percentage of the expected fare, but this is not significant ($p = 0.15$, Robust Rank Order Test).

Result 6 could be explained by drivers' beliefs about their expected payoffs from their future interaction with the passenger. Interestingly, drivers give business cards uniformly across all tester ethnicities ($p > 0.1$ in all comparisons, Fisher's Exact Tests), suggesting no taste–based discrimination in who the drivers are willing to do business with.[20] Result 6 could be a consequence of statistical discrimination. There are 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.

[20]Note, that taste–based discrimination in who to do business with is very different from taste–based discrimination in granting favours, the main focus of this paper.

5.3 Structural models

The reduced form estimates given in Section 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.

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 journey, $g(x) \cdot p$, where $g(x)$ is the distance from the rank to the drop off location of the passenger 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 testers are not found to be significant determinants of amounts given at the 5% level, as outlined in Table 11, we exclude these from the structural model.[21] Implicit in our analysis is the assumption that drivers are making their decision in isolation, or that they are 'narrowly bracketing' their decisions, and thus, are not taking into account their own annual income, or the income of the passenger (Read *et al.*, 1999).

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.[22]

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).[23] 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.

Table 13 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.

[21] 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).

[22] This model of taxi is chosen as it is the most common amongst the drivers we surveyed, as shown in Table 17 in Section 7. In reality, there are only small differences in fuel efficiency between models.

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

Model	Functional Form	Description
(1)	$u(y,m) = my^{\theta}$	Cobb–Douglas
(2)	$u(y,m) = m + \theta y^{\alpha}$	Altruistic[a]
(3)	$u(y,m) = (m^{\alpha} + \theta y^{\alpha})\alpha^{-1}$	CES[b]

[a] When $\alpha = 1$, Models (2) and (3) are identical.
[b] Constant Elasticity of Substitution.

Table 13: Estimated Functional Forms

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 an altruistic utility function is selected in Model (2): incorporating a convexity parameter will allow us to examine if utility is linear in own and others' payoffs, as is often assumed. 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 B, and the estimates presented in Table 14.

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,

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

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 e, therefore capture the additional effects of variations in the drivers' and passengers' ethnicity on θ.

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 14 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 14 under the *With Identity* heading. The parameters are estimated using only the journeys from the *Baseline* treatment to avoid any potential confounding effects originating from the *Business Card* treatment, with observations clustered at the tester level.

Models (1), (2) and (3) in Table 14 each outline how the drivers have other–regarding preferences. In the single parameter specification of Model (1), although θ is not significantly different to 0 ($p > 0.1$), the dispersion of preferences, σ, is found to be large and significant, suggesting many of the drivers do have other–regarding preferences. In Models

		Model					
		Without Identity			*With Identity*		
Parameter		(1)	(2)	(3)	(1)	(2)	(3)
Other–Regard	$\bar{\theta}$	0.045	0.895***	0.746***	0.249***	1.455***	0.803***
		(0.053)	(0.145)	(0.07)	(0.065)	(0.259)	(0.071)
Convexity	α		0.600***	0.799***		0.622***	0.829***
			(0.098)	(0.035)		(0.103)	(0.032)
Standard Deviation	σ	0.517***	1.201**	0.273***	0.454***	0.994**	0.218***
		(0.127)	(0.481)	(0.042)	(0.132)	(0.408)	(0.039)
Driver–Passenger Interactions							
White–Black	a				-0.689***	-0.276	-0.124
					(0.237)	(0.192)	(0.089)
White–S.Asian	b				0.451	0.129	0.179
					(0.397)	(0.127)	(0.173)
S.Asian–White	c				-0.829***	-0.405**	-0.039
					(0.289)	(0.190)	(0.030)
S.Asian–Black	d				-2.23***	-0.978***	-0.052
					(0.607)	(0.273)	(0.079)
S.Asian–S.Asian	e				-0.940	-0.42	0.085
					(0.888)	(0.407)	(0.085)
Observations		143	143	143	132	132	132
Log–Likelihood		-307.9053	-287.7556	-222.5028	-274.1905	-255.3374	-194.8990

No Identity: Vuong Selection Tests			
Null Hypothesis[a]	*Test Statistic*	*p–value*	*Best Fit*
H_0 : Model (1) $\leq$ Model (2)	-2.18	$p = 0.011$	Model (2)
H_0 : Model (1) $\leq$ Model (3)	-8.91	$p < 0.001$	Model (3)
H_0 : Model (2) $\leq$ Model (3)	-8.32	$p < 0.001$	Model (3)

With vs. Without Identity LR Tests[b]			
Null Hypothesis	χ^2 *Statistic*	*p–value*	*Best Fit*
H_0 : Model (1), *Without*=*With Identity*	67.43	$p < 0.001$	*With Identity*
H_0 : Model (2), *Without*=*With Identity*	64.84	$p < 0.001$	*With Identity*
H_0 : Model (3), *Without*=*With Identity*	55.21	$p < 0.001$	*With Identity*

Note: Robust standard errors clustered at the Tester level in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. Only journeys from the *Baseline* treatment are used. Journeys where the driver stopped before the meter reached the amount the tester could afford are coded as the driver giving £0. Three *Short* distance journeys in the 99th percentile of the distribution, where giving was greater than £2.50, are truncated to £2.50 due to (potentially) long wait times. *With Identity* models exclude observations from black drivers due to a lack of observations.

[a] The Vuong Test null hypothesis is that the expected likelihood contribution for each observation to Model (i) and Model (j) is equal, i.e. that both models are equally close to the 'true' model. A positive (negative) test statistic that is significant implies Model (i) (Model (j)) is preferable. See Wooldridge (2010, p.506) for a detailed discussion and Appendix B for the procedure.

[b] Likelihood Ratio Tests.

Table 14: Structural Parameter Estimates

(2) and (3), θ is estimated to be positive and significant at the 1% level, with significant preference heterogeneity reported, with $\sigma > 0$ ($p \leq 0.01$ across models). As a comparison to similarly estimated parameters in the literature, Cox *et al.* (2007) estimate Model (3) using dictator game data, with slightly different assumptions regarding the distribution of θ, and report $\theta = 0.417$ and $\alpha = 0.255$.

In the models *With Identity*, where the identity parameters are unrestricted, a number of patterns emerge. First, the estimates of α, $\bar{\theta}$ and σ remain robust. Second, estimates of a and d are always estimated to be negative, suggesting that white and South–Asian drivers' other-regarding preferences are significantly smaller when faced with a black passenger, in comparison to both white and South–Asian passengers. This is consistent with the reduced form estimates. Third, c is estimated to be negative across models, suggesting South–Asian drivers' treat white passengers worse than do white drivers. Interestingly, Model (3) estimates all identity parameters to not be significantly different to zero ($p > 0.1$ in all cases).

To help select between the competing functional forms, and examine the identity parameter restrictions, Table 14 presents the Vuong model selection statistics (Vuong, 1989) to compare the non–nested *Without Identity* models, and also Likelihood Ratio Tests to test the restrictions on the *With Identity* models.[24] As Table 14 shows, when comparing models *Without Identity*, both Model (2) and Model (3) are preferable to Model (1) ($p = 0.011$, $p < 0.001$, one sided) whilst Model (3) provides a better fit than Model (2) ($p < 0.001$, one sided). When examining restrictions on the identity parameters, the Likelihood Ratio Tests reveal that all *With Identity* models perform significantly better than the corresponding *Without Identity* models ($p < 0.001$ in all cases). Taking the results of the Vuong and Likelihood Ratio Tests together, Model (3) *With Identity* is determined to be the most preferable.

We now compare the behaviour that each functional form predicts. First, we use the estimated parameters to predict the taxi drivers' behaviour when asked to give some of the journey for free, for both *Short* and *Long* distance journeys. Second, as the taxi drivers face a situation that is analogous to laboratory dictator games, we use the preference estimates to predict their behaviour in a £10 dictator game. Doing so serves two purposes. First, it allows us to examine the robustness of the conclusions drawn from the Vuong selection Tests and Likelihood Ratio Tests. Second, we can benchmark the predictions our estimates produce against the behaviour observed in the literature. Table 15 presents the utility maximising decision of a driver in each situation, for each of the functional forms.

As shown in Table 15, Model (3) produces good predictions of driver giving, with predicted behaviour close to the observed within sample mean. In comparison, the within sample predictions of Models (1) and (2) are poor. This is in line with the conclusions drawn Vuong statistics in Table 14, that suggest Model (3) provides the best fit for the data. Table 15 also shows how the dictator giving prediction of Model (3) is reasonably consistent with that observed in the literature, whereas Models (1) and (2) drastically under–predict dictator giving. We take this as a sign that the estimates of Model (3) are consistent with the other–regarding preference literature.

[24] We follow the procedure of Wooldridge (2010, p.505-507) in order to calculate the Vuong test statistics.

Without Identity	*Taxi Giving*		*Dictator Game*
Model	*Short* distance	*Long* distance	
(1)	£0.16	£0.33	£0.43
(2)	£0.22	£0.22	£0.21
(3)	£0.69	£1.39	£1.89
Observed:			
Mean	£0.56	£1.11	£2.84[a]
Median	£0.40	£0.76	

[a] Mean dictator giving is taken from the meta-analysis of Engel (2011), which is reported to be 28.35% of the dictator endowment.

Table 15: Structural Model Behavioural Predictions

6 Robustness checks

6.1 Stopping Distances

Results 1 and 2 are examined for robustness by assuming that *some* giving behaviour is an artefact of the drivers finding a convenient location for the passenger to alight. Examining Result 1 first, we assume that the driver requires either 1, 2 or 3 additional charges on the meter (approximately 190, 380 and 570 yards respectively) in order to find a convenient location to stop.[25] In doing so, giving is now defined as the driver continuing past the amount the tester can afford *minus* the assumed cost associated with the stopping distance. Table 16 presents average giving under the three different assumptions about stopping distance using observations from the *Baseline*, *Short* and *Long* distance treatments.

Assumed Stopping Distance	Amount Given (£) *Short*	*Long*	% Giving More than £0.
190 Yards	0.459*** (0.68)	1.028*** (1.329)	71.3%
380 Yards	0.347*** (0.626)	0.896*** (1.275)	58%
570 Yards	0.266*** (0.564)	0.787*** (1.208)	42.7%

Note: Standard deviations in parentheses. *** denotes significance at the 1% level. Amounts given only include *Baseline* observations. When giving defined in this way produces a negative amount given, we assume the driver decided to give nothing.

Table 16: Robustness Checks on Driver Giving

As can be seen in Table 16, giving is still significantly different to zero for both *Short* and *Long* distance journeys, regardless of the assumed stopping distance. Drivers also give significantly more in the *Long* distance treatment in comparison to the *Short* distance

[25] See Table 1 for the exact amounts.

treatment under all three stopping distance assumptions ($p = 0.036$, $p = 0.037$, $p = 0.037$, Robust Rank Order Tests). Further, the majority of drivers still give more than zero, except when a conservative stopping distance of 570 yards is assumed. Even then, almost half of all drivers still give positive amounts. As such, we conclude that Result 1 is robust to stopping distances.

To examine the robustness of Result 2, we calculate the amount given as a percentage of the expected fare assuming that drivers require short stopping distances for *Short* distance journeys (approximately 190 yards) but longer stopping distances for *Long* distance journeys (approximately 570 yards). It may be that drivers use different types of roads in order to complete different length journeys, resulting in different stopping distances. Under this assumption, amounts given as a percentage of the expected fare are 7.4% in the *Short* distance journeys, and 7.6% in the *Long* distance journeys, with these percentages not being significantly different from one another ($p = 0.465$, Robust Rank Order Test). As with Result 1, we conclude that Result 2 is robust to potential stopping distance confounds.

6.2 Multiple Hypothesis Testing

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 5, Table 22, given in Appendix B, 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, or between, treatments.

Table 21, also given in Appendix B, 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 11 and Table 12 in Section 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 Tables 22 and 21 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 (Table 22 and Table 21). The difference in giving between South–Asian and black testers is not as robust, and only remains significant when all observations from both the *Short* and *Long* distance treatments are pooled (Hypothesis 9, Table 22). Second, although the non–parametric results in support of Result 5 are not found to be significant once adjusted (Hypotheses 10–12, Table 22), the parametric results are found to be robust (Hypotheses 19–21, Table 21). However, Result 6 does not appear to be as robust as Results

4 and 5 (Table 22).

7 Discussion

Three main questions arise from the results in Section 5: (1) can the extent of giving be explained by the drivers finding a convenient location to stop?; (2) can the drivers' behaviour be explained by their expectations relating to passengers bargaining?; (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.[26] 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 a sample *Short* and *Long* distance journey, where the sample 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.[27] 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 17, Panel A, and the passenger responses and the recorded observation results are presented in Panel B. The responses in Table 17 highlight three main points relating to (1). First, the vast majority of taxi journeys are taken from ranks (92%). Second, the majority report returning to a home, busy city rank (73%). This suggests that giving to the passenger, by continuing to drive away from the rank, is not done at the drivers' convenience. On the contrary, driving away from the rank is the same as driving away from the next passenger, and therefore is costly. Third, drivers are given ample time to stop in both the *Short*, and *Long* distance treatments, and there is no reason to think the distance required to find a convenient location to stop is proportional to the length of the journey.

Addressing (2), we note from Table 17 that the vast majority of drivers said they would not bargain with a passenger before the passenger was inside the vehicle for the *Short* (*Long*) distance journey 96% (88%). The lowest fare they would accept is also above the amount our testers could afford. In addition, the majority of drivers would refuse to bargain with passengers mid–journey (96%). Drivers' expected fare estimates are also in–line with our own calculations, suggesting they can accurately calculate how much each journey will cost. Our survey and observation of passengers also show 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

[26] The survey and observations were conducted in Manchester. The questionnaire is given in Appendix A.

[27] 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)
	% of journeys that start at a rank	92%
	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$^{\diamond}$	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$^{\diamond}$	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.
$^{\diamond}$ The majority of drivers specifying this response outlined that they would return to different rank in the centre of the city.

Table 17: Driver and Passenger Survey Responses

drivers are accustomed to bargaining with passengers as the vast majority of journeys are not bargained over.[28]

[28]Passengers are allowed to bargain with drivers *ex–ante*, or before they enter the taxi. However, if no

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.

8 Conclusion

We report evidence that the majority of taxi drivers express other–regarding preferences in a competitive market setting, with limited possibilities for repeated interactions. This evidence suggests that taxi drivers have well defined preferences over relative payoffs, a finding that resonates with the results of numerous laboratory experiments and behavioural theories of social preferences. We show that our findings are robust to a wide range of controls and a variety of potential behavioural and statistical confounds. In addition, we estimate preference parameters structurally, which we then benchmark against behaviour observed in the dictator game literature. We find that the other–regarding preference parameters we estimate are reasonably consistent with the existing literature, and make good predictions both within and out of sample.

Variation in the ethnicity of the driver and the tester also allows us to explore recent theories of taste–based 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. It is also robust to correcting for potential multiple comparison problems. 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.

Our results contribute to the debate surrounding the external validity of laboratory experiments that measure other–regarding preferences, and should appease critics that are sceptical of the generalisability and interpretability of such experiments. However, 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. Especially in markets where bilateral face to face interactions are common place we expect other–regarding preferences to play a much greater role than previously suggested.

discount is agreed prior to the journey beginning, the driver is allowed to charge the metered fare by law.

References

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, Working Paper 22113.

Ayres, I., Banaji, M. & Jolls, C. (2015), 'Race effects on ebay', *The RAND Journal of Economics* **46**(4), 891–917.

Ayres, I. & Siegelman, P. (1995), 'Race and gender discrimination in bargaining for a new car', *American Economic Review* **85**(3), 304–321.

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.

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.

Bertrand, M. & Duflo, E. (2016), Field experiments on discrimination. National Bureau of Economic Research, Working Paper 22014.

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.

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.

Camerer, C. (2015), The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List, *in* G. R. Frèchette & A. Schotter, eds, 'Handbook of Experimental Economic Methodology', Oxford University Press.

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.

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.

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.

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.

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

DellaVigna, S., List, J. A. & Malmendier, U. (2012), 'Testing for Altruism and Social Pressure in Charitable Giving', *Quarterly Journal of Economics* **127**(1), 1–56.

Doleac, J. L. & Stein, L. C. (2013), 'The visible hand: Race and online market outcomes', *The Economic Journal* **123**(572), F469–F492.

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.

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.

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

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. National Bureau of Economic Research, Working Paper 17855.

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.

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*.

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.

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

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.

Kube, S., Maréchal, M. A. & Puppea, 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.

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

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.

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. (2014), 'Are good-looking people more employable?', *Management Science* **61**(8), 1760–1776.

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.

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.

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.

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.

Vuong, Q. H. (1989), 'Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses', *Econometrica* **57**(2), 307–333.

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

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

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.

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.

A Experimental Appendix

A.1 Job advertisement

[redacted]

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 [redacted] Applicants must have the right to work in the UK. Proof of this will be required if you are successful.

Email:

A.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**

A.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	☐		

A.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.

A.5 Ex-post driver survey

[redacted]

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

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?**

B Statistical Appendix

B.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

$$u(x; \alpha, \theta, a, b, c, d, e, \epsilon_x) = u(x+1; \alpha, \theta, a, b, c, d, e, \epsilon_x). \tag{2}$$

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 2 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,

$$\frac{\epsilon_x}{\sigma} = \frac{1}{\sigma}\left(\frac{(s - x - g(x) \cdot p)^\alpha - (s - x - g(x+1) \cdot p - 1)^\alpha}{(x+1)^\alpha - x^\alpha} - \theta\right). \tag{3}$$

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

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

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

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

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

$$\Pr[x^* = \bar{x} | \alpha, \theta, a, b, c, d, e, \sigma] = \int_{\epsilon_{\bar{x}-1}}^{\infty} f(z)dz = 1 - F(\epsilon_{\bar{x}-1}). \tag{6}$$

The likelihood function, using 132 journeys from the *Baseline* treatment, as specified in Section 5.3, is therefore

$$L(\alpha, \theta, a, b, c, d, e, \sigma) = \prod_{k=1}^{132} Pr[x_k = x; \alpha, \theta, a, b, c, d, e, \sigma]. \tag{7}$$

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

B.2 Vuong Test Statistics

The Vuong model selection test statistics are calculated following Wooldridge (2010). The null hypothesis for each test is

$$H_0 : E\left[\ell_{1i}(\boldsymbol{\beta}_1^*)\right] \leq E\left[\ell_{2i}(\boldsymbol{\beta}_2^*)\right] \tag{8}$$

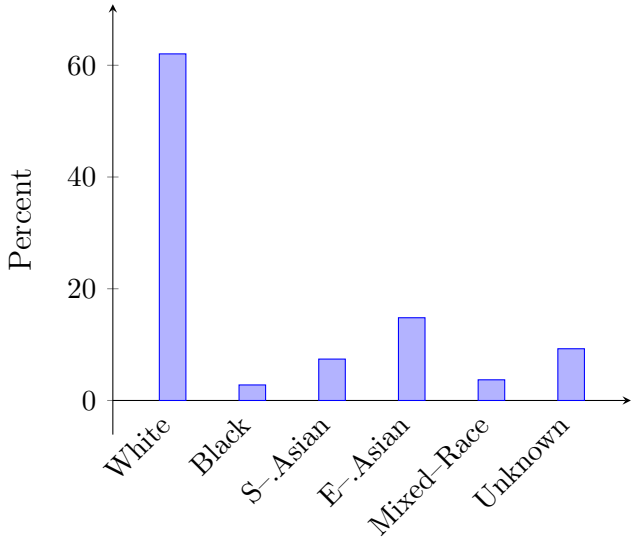
where $\ell_{ki}(\boldsymbol{\beta}_k^*)$ denotes the likelihood contribution of observation i to model $k \in \{1, 2\}$, where $\boldsymbol{\beta}_k^*$ are the parameters that model k converges to. Wooldridge (2010) outlines how a valid test statistic can be obtained by regressing the difference between the estimated individual log likelihood contributions of each observation from each model specification, $\hat{d}_i = \ell_{1i}(\hat{\boldsymbol{\beta}}_1) - \ell_{2i}(\hat{\boldsymbol{\beta}}_2)$, on unity and then testing if the coefficient is significantly different from zero. The results from these regressions, used for comparing the models in Table 14, are given in Table 18.

Model Comparison	Coefficient	Robust Std. Error	Test Statistic	p–value
(1) vs (2)	-0.141	0.065	-2.18	0.016
(1) vs (3)	-0.597	0.067	-8.91	0.000
(2) vs (3)	-0.456	0.055	-8.32	0.000

Note: p–values are one sided.

Table 18: Calculated Vuong Test Statistics

B.3 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 4: Distribution of the Raters' Self–Reported Ethnicity

B.4 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 19: Tester Appearance Correlations

B.5 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 4.1. Observations with a white driver and a white Tester are taken as the baseline.

Table 20: The Effects of Ethnic Interactions on Giving

B.6 Robustness checks

Table	#	*Result*	*Model*	*Explanatory Variable of Interest*			
				Black	*South–Asian*	*Male*	*m*
	1		(1)	0.002**	0.389		2
	2		(2)	0.015**	0.274	0.015**	3
Table 11	3	*Result 4*	(3)	0.013**	0.23	0.017**	3
	4		(4)	0.026**	0.276	0.034**	3
	5		(5)	0.069*	1.00	0.121	8
	1		(1)	0.025**	0.195		2
	2		(2)	0.038**	0.116	0.233	3
Table 12	3	*Result 5*	(3)	0.027**	0.093*	0.248	3
	4		(4)	0.043**	0.106	0.313	3
	5		(5)	0.065*	0.704	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 21: Adjusted p–values – Parametric Testing

#	*Alt. Hypothesis*	*Family*	*Outcome*	*Unadjusted*	*Adjusted*
Result 4					
1	H_A: White ≠ Black	*Baseline, Short*	Giving, £	0.001***	0.003***
2	H_A: White ≠ S.–Asian			0.41	0.41
3	H_A: S.–Asian ≠ Black			0.069*	0.138
4	H_A: White ≠ Black	*Baseline, Long*	Giving, £	0.374	1.00
5	H_A: White ≠ S.–Asian			0.88	0.88
6	H_A: S.–Asian ≠ Black			0.47	0.47
7	H_A: White ≠ Black	*Baseline, pooled*	Giving, %	0.004***	0.008**
8	H_A: White ≠ S.–Asian			0.361	0.361
9	H_A: S.–Asian ≠ Black			0.028**	0.056*
10	H_A: White ≠ Black	*Business Card, 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	*Business Card, Long*	Giving, £	0.566	0.566
14	H_A: White ≠ S.–Asian			0.07*	0.14
15	H_A: S.–Asian ≠ Black			0.45	0.90
16	H_A: White ≠ Black	*Business Card, pooled*	Giving, %	0.0005***	0.0015***
17	H_A: White ≠ S.–Asian			0.003***	0.006***
18	H_A: S.–Asian ≠ Black			0.76	0.76
Result 5					
19	H_A: White ≠ Black	*Baseline, 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: *B.line* ≠ *B.card*, White	*B.line vs B.card, pooled*	Giving, %	0.11	0.33
23	H_A: *B.line* ≠ *B.card*, Black			0.43	0.43
24	H_A: *B.line* ≠ *B.card*, S.–Asian			0.19	0.29

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 22: Adjusted p-values – Non–Parametric Testing