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

Brit Grosskopf Graeme Pearce

University of Exeter

May 17, 2016

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–subjects design we vary the length of the journey and whether drivers have reputational concerns or not. We find that the majority of drivers give at least part of the journey for free and over 25% complete the journey. Giving is found to be proportional to the length of the journey, and the drivers' reputational concerns do not explain their behaviour. Evidence of strong out–group negativity against black testers by both white and South–Asian drivers is also reported. In order to link our empirical analysis to behavioural theory we 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.

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

*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 Henry Schneider for comments. We thank the University of Exeter Business School for funding this research.

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

Other–regarding preferences also form the foundation for recent behavioural theories of discrimination. Stemming from concepts of 'taste–based' discrimination first detailed in Becker (1971), a prominent theory is that social preferences are group–contingent, or that other–regarding preferences are larger towards those we identify with (the 'in–group'), in comparison to 'out–groups' (Chen & Li, 2009). Although this explanation has gained prominence, as with other work on social preferences, the majority of evidence in its support has been obtained from laboratory experiments (Chen & Chen, 2011; Drouvelis & Nosenzo, 2013; Goette *et al.*, 2006; van Der Mewe & Burns, 2008). Field experiments, in contrast, largely suggest discriminatory behaviour can be attributed to statistical discrimination (List, 2004; Levitt, 2004; Gneezy *et al.*, 2012), although some come close to identifying a taste (Bertrand & Mullainathan, 2004; Mujcic & Frijters, 2013). In addition, the methods used for studying identity and discrimination in the laboratory have recently been criticised, with work suggesting the observed behaviour is a consequence of experimenter demand effects (Zizzo, 2010, 2012), or possibly stemming from a heuristic (Guala & Filippin, 2015), rather than being due to an inherent preference.

The purpose of this paper is to examine the prevalence, and extent, of other–regarding preferences in a highly competitive market place, and determine the role played by reputational concerns in fostering other–regarding behaviour. We also investigate the significance of ethnic identity in determining these preferences, and examine its interplay with individuals' reputational concerns. This is done using a natural field experiment whereby we employed 22 testers of varying ethnicity to pose as passengers and take a number of pre–determined taxi journeys.[1] In each case we endowed them with only 80% of the expected fare. Once the taxi meter reached 60% of the fare, testers told the driver that they only had a certain amount, and asked if they could have the final 20% of the journey for free. The tradeoff faced by a driver in this situation is analogous to the dilemmas that subjects typically face in the laboratory: express other–regard at a personal cost but to the benefit of another by giving some of the journey for free, or to behave selfishly but profitably by stopping once the meter reaches the amount the passenger can afford.[2]

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

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

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

ment are able to give twice as much (in absolute terms) as drivers assigned to the *Short* distance treatment, we can examine if the drivers' other–regarding preferences depend on the relative payoffs between themselves and the passenger, or if giving is constant regardless of the amount available to give. Orthogonally to the distance treatments, we vary whether the drivers' reputations are a concern to them or not. Using a *No Reputation* treatment, testers signal the one–shot nature of the interaction to the driver. The taxi markets we study have thousands of drivers, and tens of thousands of passengers each week, making repeated interactions for infrequent customers incredibly unlikely; these markets are therefore attractive for studying the 'one–shot' interactions required for disentangling other–regard from reputational concerns. As described in Section 3, the only real possibility of meeting a driver in a future interaction is by obtaining his contact details so that he can be actively selected. Our *Reputation* treatment, similar to the repeat business treatment of Schneider (2012), exploits this with testers asking drivers for a business card so they can contact them for future journeys. Making drivers' reputations salient will allow us to examine how the prospect of a repeated interaction affects drivers' other–regarding behaviour.

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

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

In the *Reputation* treatment we find that reputational concerns can increase the drivers' other–regard, but only when drivers are carrying a white tester. Black testers see no significant increases, and we observe decreases in giving for South–Asian testers stemming from reputational concerns. The differential effect of reputation is attributed to the drivers' beliefs being influenced by the passengers' ethnicity, either their belief about the probability of a repeated interaction, or their belief regarding the payoff they will receive from the future interaction. This is discussed further in Section 5.

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

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

The remainder of this paper is organised as follows. Section 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 by accounting for potential multiple hypothesis testing. 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), other–regarding preferences are well established to exist in the laboratory. However, it is typically assumed that social preferences are irrelevant in market settings (Schmidt, 2011) and, as reported by DellaVigna (2009), there is little field evidence to support many of the laboratory derived conclusions. Levitt & List (2007) provide a range of reasons for why these conclusions may fail to generalise to field settings. In addition, pro–social behaviour in the field is often difficult to attribute to inherent preferences, as it is easily attributed to reputational concerns and social pressure effects (Akerlof & Kranton, 2000).[3]

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

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

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

from those with standard preferences.[5]

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

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

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

In the field, DellaVigna *et al.* (2012) use a novel natural field experiment, nested within a charitable door–to–door fund raiser, in order to disentangle altruism from social pressure effects. The experiment gives potential donors the option to opt–out of meeting a fund raiser, allowing those who might give as a consequence of social pressure to select out. Although they find that a significant number of individuals give out of pure preference to do so, social pressure is found to increase giving substantially. There is also evidence to suggest social pressure influences voter turn–out (Gerber

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

et al., 2008), and causes workers to partially internalise the negative externalities of free riding; Mas & Moretti (2009) find that worker effort is positively related to the productivity of workers who observe them, but also with those they expect to interact with again.

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

2.2 Discrimination

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

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

In order to try and understand discriminatory behaviour in the absence of stereotypes and beliefs that may otherwise affect behaviour, laboratory experimenters have turned to study artificially induced identities by using minimal (Tajfel *et al.*, 1971),

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

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

Those studies which are most related to ours, audit studies, utilise actors to take part in standardised interactions such as job interviews.[6] 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 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

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

against foreigners, such behaviour is consistent with drivers exploiting informational asymmetries between passengers, as non–locals are unlikely to be familiar with the average fare of a particular journey. Using observational data, Jackson & Schneider (2011) detail how New York City taxi drivers who lease a car from a member of their country–of–birth exhibit reduced effects of moral hazard. They argue that such a result is consistent with the presence of increased social sanctions in the form of community–enforced punishments, rather than being a consequence of social preferences.

Whilst not a study of taxis, the study closest to ours is that of Mujcic & Frijters (2013). Exploiting a natural interaction between bus drivers and passengers, paid testers acting as passengers attempted to board buses without any money. They find that white testers are allowed to embark 72% of the time, Indians 51% and blacks just 36% of the time. This result remains robust to a wide range of controls, including tester characteristics, such as aggression, attractiveness, and others, elicited through a post–experimental survey. These controls, which are neither elicited in an incentive compatible nor in an anonymous manner, are included in an attempt to control for the 'Heckman criticism' (Heckman, 1998): implicit in the assumptions of all audit studies is that *unobservable* characteristics of confederates are identical across gender or ethnicity. The interaction can be viewed as an other–other allocation game (Tajfel *et al.*, 1971; Turner, 1978), where the driver must allocate resources between the passenger and the bus company, rather than being comparable to the dictator game. As drivers are not monitored, their choices, while costly to the bus company, are financially costless to them. Our study distinguishes itself from Mujcic & Frijters (2013) as we consider discrimination in a situation where pro–social behaviour is costly to the person exhibiting it.

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

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

		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
Cost of 1.8 Mile Journey	£5.44	£5.26	£4.87	£4.86
Cost of 4.4 Mile Journey	£10.12	£10.16	£10.61	£9.17
Wait Time Rate (per hour)	£16.00	£18.46	£16.80	£8.00

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

Table 1: Taxi Fares by Local Authority

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.[7] All vehicles are fitted with a taxi meter which displays the cost of the journey, up to a given point, to the passenger. The meter starts from a fixed amount and increases by a set amount every so many yards driven, or seconds waiting in traffic. Metered fares are set by the local authority. Those relevant for this study are detailed in Table 1.

Important to our study is the fact that the metered fare is the maximum fare the driver is able to charge the passenger unless a different fare was negotiated prior to the passenger entering the taxi. If no ex-ante negotiation took place, the metered fare is the amount the passenger must pay by law. Where no negotiation took place, fare reductions are made entirely at the driver's discretion and the driver is within his

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

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

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

Table 2: Taxis, Taxi Ranks and Weekly Passenger Numbers

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.[8] 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. It was also designed to examine if reputational concerns can explain other-regarding behaviour. We use a natural field experiment that allows us to observe behaviour in a market setting, in a natural interaction devoid of experimenter scrutiny. Our subjects, the taxi drivers, were oblivious to a study taking place.

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

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).

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.[9] They were told the script may vary, and that they would be given a chance to practice any variants before completing the task. Testers were told explicitly to follow the script as closely as possible, and when interacting with the drivers they were told they must not attempt to influence any of their decisions. Testers were told not to engage in conversation with the drivers, and scripted responses were given to anticipated questions. Our hypotheses and predictions regarding the study were never made clear to the testers, and not all the testers met each other, reducing the opportunity for testers to guess the study might involve their own ethnicity.[10] 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

[9] We discussed the *Short* distance / *No Reputation* treatment, which is described in Section 4.2.

[10] 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.

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

Note: Testers' age and ethnicity is self-reported. Correlations between appearance characteristics are presented in Table 15 in Appendix B. The raters' ethnicities are presented in Figure 5 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

passenger regarding how much money they have, we also elicit the passengers' facial appearance of trustworthiness.

To obtain the ratings, each laboratory subject was shown a random set of 11 photos and asked to rate their appearance. Following Xiao & Houser (2005), to increase subjects' attentiveness to the task they were told that one photo, and one characteristic of that photo, would be selected at random, and if their decision for that photo and that characteristics was in line with the ratings of the majority of the other subjects in the session, they would receive £2. It took subjects around 10 minutes to rate all the photos required of them. A sample of 1188 ratings was obtained from 108 laboratory subjects. The ratings are presented in Table 3.[11,12]

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

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

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

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 £4 or £8 for each journey, depending on its distance. Journeys were taken in either Birmingham or the Greater Manchester area, with those starting in Birmingham taken over 5 days, and those in Manchester over 3. All journeys were taken between 11am and 5pm and at least 4 testers were in the field at any given time, along with an experimenter.

Upon entering the taxi, the tester first stated their destination, and then spoke a simple entry statement.[13] In the *No Reputation* treatment they stated, "I don't take taxis very often", and in the *Reputation* treatment they stated, "I'm looking for a reliable driver for future journeys. Can I have a business card?". The first statement signals to the driver that the interaction is one-shot, as a passenger who doesn't take taxis very often is unlikely to meet the same driver twice. The second statement is designed to signal that a repeated interaction is possible, that the drivers' behaviour may influence the probability of a future interaction, and may affect the payoffs from a future interaction.[14] 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

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

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

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

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

Table 4: Experimental Design Summary

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.

Driver Characteristics					
		Driver Ethnicity			
	All Drivers	White	Black	South–Asian	Other
Age	44.3	50.06	40.36	42.6	41.3
	(10.67)	(10.56)	(9.36)	(10.03)	(11.45)
Gender (1 if *male*)	0.99	0.97	1	0.99	1
	(0.12)	(0.17)	(0)	(0.07)	(0)
Journeys	283	71	11	191	10
Field Characteristics					
			Mean		
Traffic (1 if Not Busy, 10 Very Busy)			4.44		
			(2.26)		
Weather (1 if *raining*)			0.11		
			(0.32)		
Ride Characteristics					
			Mean		
Conversation (1 if driver attempted a *conversation*)			0.28		
			(0.45)		
Cashpoint (1 if driver offered a *cashpoint*)			0.04		
			(0.2)		
Business card, *Reputation* only (1 if one was given)			0.45		
			(0.5)		
Receipt Given (1 if given)			0.9		
			(0.308)		

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

Table 5: Variables Recorded by the Testers

We refer to the driver continuing the journey past the amount that the tester can

afford as *giving*, or as the driver expressing his other–regarding preferences, which is accurately measured by the meter. Once the driver decided how much to give, and where to end the journey, the tester had to ask for a receipt, leave the taxi, and discretely complete an experimental sheet. The sheet included subjective characteristics of the driver, such as his age, gender (1 if male) and ethnicity, measures of the field including traffic intensity (recorded on a 10 point scale: 1 if Not Busy, 10 if Very Busy) and the weather (1 if raining), and finally characteristics of the ride including whether the driver attempted a conversation (1 if yes), if he offered a cashpoint (1 if yes) and (in the *Reputation* treatment) if he gave a business card or not (1 if one was given). Most importantly, the testers had to record the final meter reading and if the driver completed the journey or not.[15] We present these measures in Table 5.

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

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 6 outlines the recorded fare, expected fare and amounts given as a percentage of the expected fare, from journeys where the driver completed the journey. Observations are disaggregated by *Short* and *Long* distance journeys. By comparing the observed fare in a completed journey to its expected fare, the accuracy of our expected fare calculations can be examined. Minor discrepancies between recorded and expected fares are to be expected, largely due to variations in traffic intensity and other random shocks.

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

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

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

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

Table 6: Fares, Expected Fares and Average Giving conditional on the Driver Completing the Journey

5.2 Other regard and reputation effects

Table 7 outlines average amounts given by drivers and the proportion of journeys they completed, by treatment. To examine if relative payoffs are a motivating factor behind the amounts that drivers are giving, giving as a percentage of the expected fare is also reported. Figure 1 displays the distribution of giving across treatments.

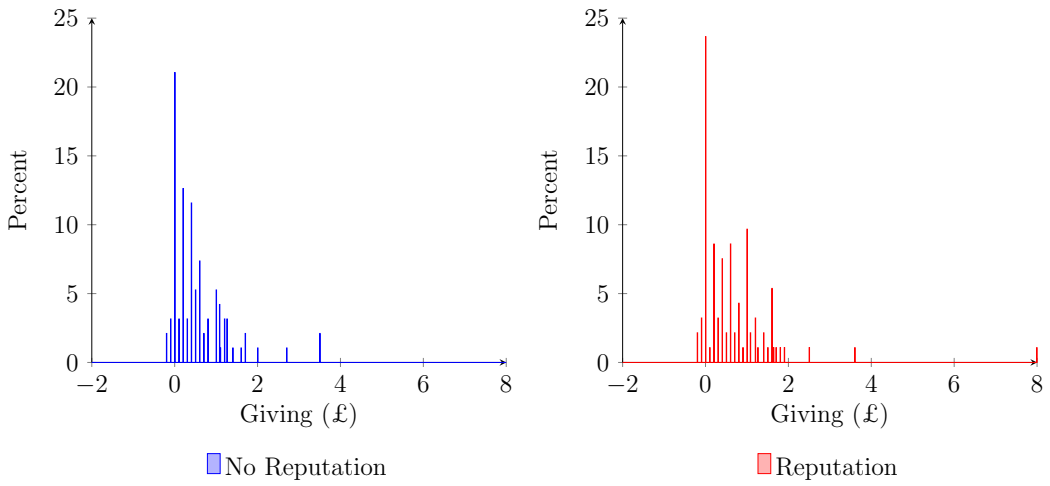
	No Reputation		*Reputation*	
	Short	*Long*	*Short*	*Long*
Amount Given (£)	£0.56	£1.11	£0.71	£1.07
	(0.69)	(1.39)	(1.06)	(1.21)
Amount Given as a % of the Expected Fare	10.6%	11.2%	13.4%	10.5%
	(0.128)	(0.144)	(0.207)	(0.12)
Proportion of Journeys Completed	0.27	0.27	0.31	0.34
Number of Journeys	95	48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

Note: Standard deviations in parentheses.

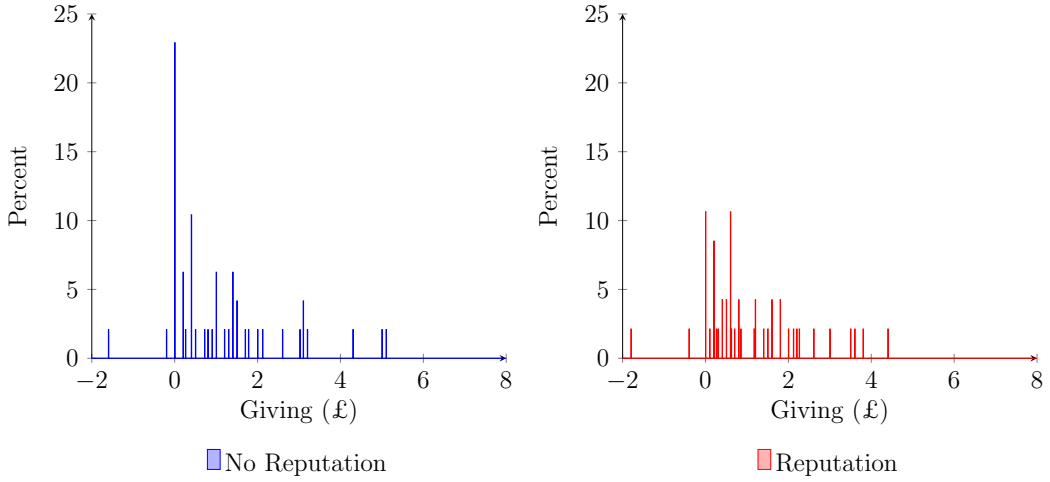
Table 7: Average Driver Giving, by Treatment

Table 8 reports a number of random effects Tobit regressions. In models (1), (2) and (3) giving in pounds by driver i to tester j is the dependent variable. In models (4), (5) and (6), giving as a percentage of the expected fare by driver i to tester j is the dependent variable. Considering giving in this way enables us to control for the variation in journey lengths, and therefore variation in the expected fares of journeys, both within and between treatments. In each regression, dummy variables for the *Long* distance treatment and the *Reputation* treatment are included along with their interaction; the *Short* distance *No Reputation* treatment is taken as the baseline.

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



(a) Short Distance Journeys



(b) Long Distance Journeys

Figure 1: Distribution of Giving by Treatment

the variable for traffic intensity (recorded on a 10 point scale: 1 if Not Busy, 10 if Very Busy), and a dummy controlling for the weather conditions (1 if raining). City Controls includes dummies for the journey taken in Birmingham, Trafford or Salford (1 if yes), with those taken in Manchester taken as the baseline. Ride Controls includes dummies controlling for whether the driver offered to take the passenger to a cash–point (1 if offered) and if he tried to engage in a conversation (1 if yes).

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

Support. Considering journeys from the *No Reputation* treatment, the null hypothesis of no giving can be rejected at the 1% level in both *Long* and *Short* distance journeys ($p < 0.01$, both cases, Sign Test). Over 70% of drivers give part of the journey for free, and over 25% of all journeys were completed in full. Parametric support is given in Table 8, with a positive and significant constant in all regression models

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

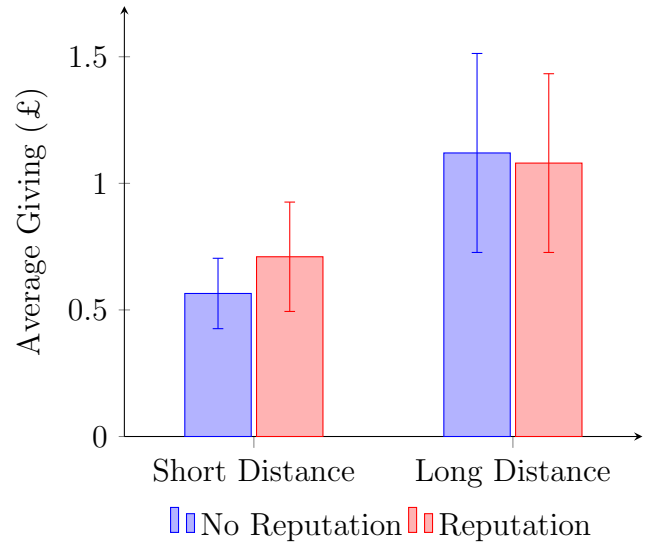
Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. The number of observations fall slightly as more controls are included due to missing entries. Models (1), (2) and (3) are left censored at 0, and right censored at the difference between the expected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

Table 8: Treatment effects

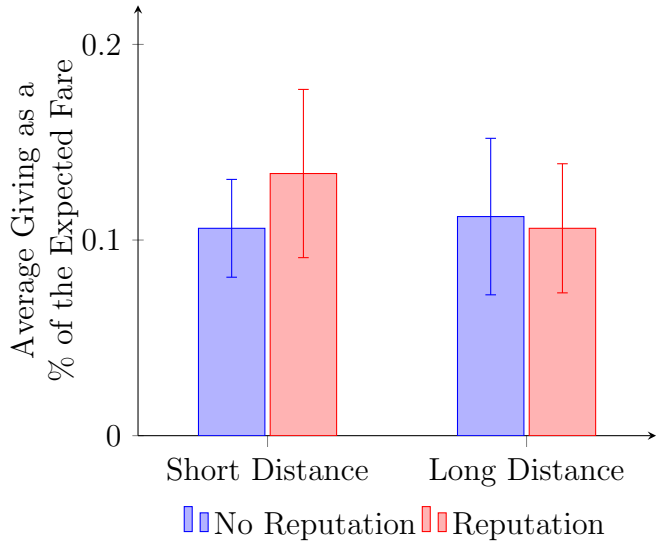
($p < 0.05$, in all cases). Similar findings are observed in the *Reputation* treatment, with over 75% of drivers giving at least part of the journey for free, and 32% of all journeys being completed in full.

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

Support. Examining journeys from the *No Reputation* treatment, average driver giving is significantly different in *Short* distance journeys in comparison to *Long* distance journeys ($p = 0.056$, Robust Rank Order Test). This is shown graphically in Figure 2a. The distribution of giving is also found to vary by the distance of the journey ($p = 0.039$, Kruskal–Wallis Test). Table 8, regressions (1), (2) and (3) support these conclusions, reporting significant and positive coefficient estimates on the *Long* distance dummy ($p < 0.05$), whilst the coefficient on the *Reputation* dummy alone is not significant ($p > 0.1$). However, when giving as a percentage of the expected fare is considered, no significant differences are reported by distance ($p = 0.86$, Robust Rank Order Test) (see Figure 2b). Further, the distance of the journey has no significant effect on its distribution ($p = 0.86$, Kruskal–Wallis Test). Estimates from Table 8 models (4), (5) and (6), support this conclusion; no significant treatment effects are reported when the dependent variable is giving as a percentage of the expected fare ($p > 0.1$ in all cases, in all regressions), suggesting giving is proportional to the length of the journey, and therefore the amount the driver can give.



(a) Average Giving (£)



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

Note: Vertical bars represent 95% confidence intervals.

Figure 2: Average Giving

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

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

Support. Comparing average giving between *Reputation* and *No Reputation* treatments, no significant differences are reported in either *Short* or *Long* distance journeys ($p = 0.34$ and $p = 0.67$, Robust Rank Order Tests). Similarly, reputational concerns have no significant impact on the distribution of giving in either *Long* or *Short* distance treatments ($p = 0.44$ and $p = 0.67$, Kruskal–Wallis Test). The same is true for giving as a percentage of the expected fare, with no significant differences found between *Reputation* and *No Reputation* treatments in *Short* or *Long* distance journeys, or when journeys are pooled ($p > 0.1$ in all cases, Robust Rank Order Tests). Estimates from Table 8 supports these results, with the coefficient on the *Reputation* dummy found to be not significant at conventional levels across regressions ($p > 0.1$ in all cases).

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

In addition, the drivers' behaviour may depend on the appearance characteristics of the tester. To explore this further we examine driver giving conditional on the testers' ethnicity. Summary statistics are given in Table 9 and Figure 3 displays the proportion of completed journeys by tester ethnicity graphically. To determine the effect of the testers' ethnicity on driver giving, Table 10 outlines the results from a number of random effects Tobit regressions. In each case, giving in pounds by driver i to tester j is the dependent variable. The estimated coefficients on a dummy controlling for whether the tester was black (1 if yes), South–Asian (1 if yes) and if they were male (1 if yes) are reported; white testers are taken as the baseline.

To examine the robustness of the estimated coefficients, in each model we systematically increase the number of explanatory variables, which are grouped into 6 sets: Treatment, Driver, Tester, Ride, Field and City Controls. Treatment controls include dummies for each of the treatments (1 if *Long*, and 1 if *Reputation*) and the interaction, Driver controls include the driver's age and gender (1 if male). Tester controls include the tester's gender (1 if male), which is reported, and also their age. Field, Ride and City controls are identical to those described for Table 8. For each tester, we also include their average rating for each appearance characteristic: aggressiveness, attractiveness, friendliness, trustworthiness and wealthiness. The estimated coefficients on these variables are included in Table 10.

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 *No–Reputation* treatment reveals no significant differences between white and South–Asian testers in the *Short* or *Long* distance treatments $p = 0.64$ and $p = 0.46$, Robust Rank Order Tests). However, significant differences

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

Note: Standard errors given in parentheses.

Table 9: Summary Statistics by Tester Ethnicity.

between white and black testers are reported in the *Short* but not in the *Long* distance treatment ($p = 0.001$ and $p = 0.39$, Robust Rank Order Tests). Similarly, a significant difference between South–Asian and black testers is found in the *Short* but not in the *Long* distance treatment ($p = 0.06$ and $p = 0.47$, Robust Rank Order Tests). Considering giving by the amount given as a percentage of the expected fare reveals that both white and South–Asian testers are given significantly more than black testers ($p = 0.005$, $p = 0.025$, Robust Rank Order Tests), but no differences are found between white and South–Asian testers ($p = 0.31$, Robust Rank Order Test). The estimates in Table 10 further support the non–parametric results: across all regressions, the coefficient on the black dummy is negative, highly significant ($p < 0.01$, Wald Tests), and robust to changes in the model specification.

The differential treatment of testers by ethnicity remains in the *Reputation* treatment, with white testers receiving more than black testers in the *Short* distance 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 $p = 0.003$, Robust Rank Order Tests), but no difference between black and South–Asian testers ($p = 0.9$, Robust Rank Order Test).

The proportion of completed journeys, by tester ethnicity, is now considered. Table 11 reports a number of random effects Probit regressions, where the dependent

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

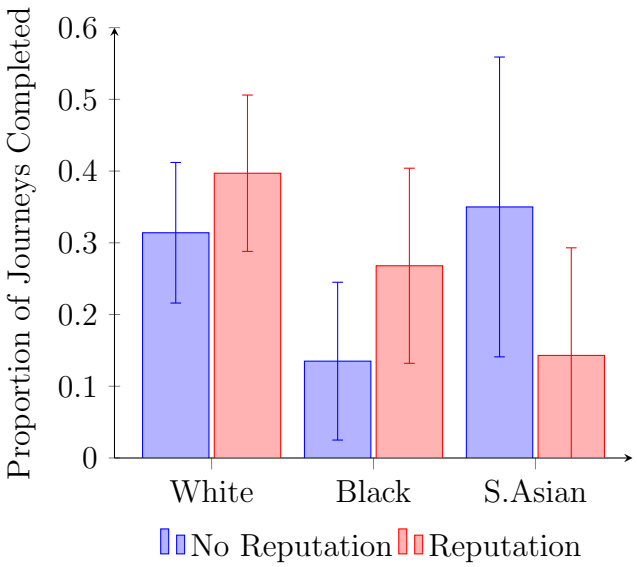
Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. The number of observations falls slightly as more controls are included due to missing entries. All models are left censored at 0, and right censored at the difference between expected fare, had the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

Table 10: The Determinants of Driver Giving

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 10.

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

Support. Comparing the proportion of journeys that were completed, by tester ethnicity, black testers have their journey completed significantly less often than white and South–Asian testers in the *No Reputation* treatment ($p = 0.045$ and $p = 0.088$, Fisher's Exact Test). No significant differences are reported between white and South–Asian testers ($p = 0.793$, Fisher's Exact Test). The results from the random effects Probit regressions in Table 11 outline how the estimated coefficient



Note: Vertical bars represent 95% confidence intervals.

Figure 3: Proportion of Journeys Completed by Tester Ethnicity

on the black dummy is negative and significant ($p = 0.05$). This estimate is robust to specification changes, and becomes increasingly significant as more controls are included. Similar to the coefficient estimates in Table 10, none of the appearance characteristics are significant, except the appearance of wealthiness ($p = 0.06$), which has a negative effect: more wealthy looking testers are less likely to have a journey completed.

Results 4 and 5 outline how black testers are treated significantly worse than white and South–Asian testers. As the coefficient on trustworthiness is insignificant, and its direction the opposite we would expect, statistical discrimination is likely not the explanation. Status is also unlikely to be a factor, as wealthiness has a negative effect on giving and our black testers are rated as appearing the *least* wealthy as outlined in Section 4.1. Indeed, the inclusion of the appearance characteristics increases the magnitude of the coefficient of the black dummy in both the Tobit and Probit regressions. The evidence points towards taste based discrimination.

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

Support. White testers are given significantly greater amounts as a percentage of the expected fare in the *Reputation* treatment compared to the *No Reputation* treatment ($p = 0.06$, Robust Rank Order Test). They also receive significantly more in absolute terms as a result of reputation in the *Short* distance treatment ($p = 0.053$, Robust Rank Order Test), although no significant difference is observed in the *Long* distance treatment ($p = 0.61$, Robust Rank Order Test). Black testers see no significant differences as a result of reputation ($p > 0.1$ in all cases, Robust Rank Order

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

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

Table 11: Determinants of Journey Completion

Tests). South–Asian testers see no effect of reputation on absolute giving in both the Short ($p = 0.27$, Robust Rank Order Test) and *Long* distance treatments ($p = 0.5$, Robust Rank Order Test), and a (weakly) negative effect is reported in giving as a percentage of the expected fare ($p = 0.15$, Robust Rank Order Test) resulting from driver reputational concerns.

Result 6 can be explained by drivers' beliefs about their expected payoffs from their future interaction with the passenger, and is unlikely to be due to beliefs that white passengers are most able to contact them: drivers give business cards uniformly across all tester ethnicities ($p > 0.1$ in all comparisons, Fisher's Exact Tests). There are, however, two different belief channels through which the disparity can occur, either through drivers' beliefs about the probability of a repeated interaction, or through their beliefs about their earnings from a repeated interaction. Drivers may believe the probability of a future interaction is greatest for a white passenger, or that

by expressing other–regard they increase this probability by more than if the tester was black or South–Asian. Alternatively, drivers may believe white passengers are more likely to reciprocate their other–regard in a future interaction through tipping, as shown by Ayres *et al.* (2005), who report that white passengers in the United States tip approximately twice as much as passengers of other ethnicities.

5.3 Structural models

The reduced form estimates provided 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 of the journey in miles, and p the price in fuel per mile travelled: $m = s - x - g(x) \cdot p$. When the driver selects $x = 0$, he stops when the meter reaches the amount the passenger can afford; $x = \bar{x}$ implies he completed the journey. The passenger's payoff is defined as being equal to the amount given to her by the driver, x, so $y = x$. As the appearance characteristics of the tester are not found to be significant determinants of amounts given at the 5% level, as outlined in Table 10, we exclude these from the structural model.[16]

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.[17]

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).[18,19] 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 12 outlines the three functional forms of utility that we estimate. Due to

[16] 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).

[17] This model of taxi is chosen as it is the most common amongst the drivers we surveyed, as shown in Table 14 in Section 7. In reality, there are only small differences in fuel efficiency between models.

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

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

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

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

Table 12: Estimated Functional Forms

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

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

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

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. 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.

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 13 under the *Without Identity* heading. Second, we remove the identity parameter restrictions, and let the model pick their values; the results are displayed in Table 13 under the *With Identity* heading. The parameters are estimated using only the journeys from the *No Reputation* treatment to avoid any potential confounding effects originating from the

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

Note: Standard errors clustered at the tester level. Robust standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level, respectively. Only journeys from the *No Reputation* treatment are used, from both the *Short* and *Long* distance treatments. Journeys where the driver stopped before the meter reached the amount the tester could afford are coded as the driver giving £0. Reduced form estimates that support these results are given in Table 16 in Appendix B.

Table 13: Structural Parameter Estimates

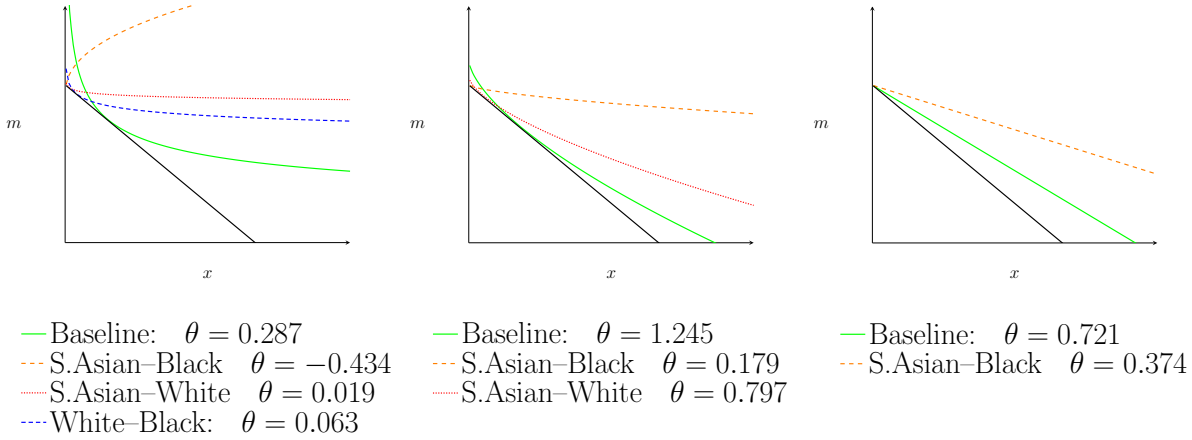
drivers' reputational concerns, with observations clustered at the tester level.[20]

Models (1), (2) and (3) in Table 13 each outline how the drivers have other–regarding preferences. In the single parameter specification of model (1), although $\bar{\theta}$ is not significantly different to 0 ($p > 0.1$), the dispersion of preferences, σ, is found to be significant, suggesting many of the drivers do have other–regarding preferences. In models (2) and (3), $\bar{\theta}$ is estimated to be positive and significant at the 1% level, with significant preference heterogeneity reported, with $\sigma > 0$ ($p \leq 0.1$ across models).[21]

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

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

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



(a) Model 1 (b) Model 2 (c) Model 3

Note: The passenger's payoff, x, is plotted on the x axis and the driver's payoff, m, is plotted on the y axis. In each panel, the thick black curve represents a hypothetical 'budget line'. In Figure 4b, $\alpha = 0.676$, and in Figure 4c, $\alpha = 0.846$, for each of the indifference curves, as estimated in Table 13. Where identity parameters are found not to be significant, other–regarding preferences are taken to be equal to the baseline.

Figure 4: Estimated indifference curves

Figure 4a).

Second, weak evidence that white driver preferences are reduced when faced with a black passengers is reported, with $a = -0.781$ ($p < 0.01$) in model (1), although its significance is not robust to specification changes. The parameter measuring the effect of South–Asian / white interactions, c, is similar, estimated to be negative and significant in models (1) and (weakly) significant in model (2) ($p = 0.06$), but insignificant in model (3) ($p > 0.1$). Finally, no evidence is found that white drivers' preferences are influenced by South–Asian passengers, with b found to be positive, but insignificant in all models ($p > 0.1$ in all cases). Figure 4 plots the estimated indifference curves from model (1), (2) and (3) graphically.

6 Robustness checks

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

For each result in Section 5, Table 17, 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 a treatment, or the group of tests related to a particular outcome compared between treatments.

Table 18, 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 10 and Table 11 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 Table 17 and 18 provide a number of insights. First, the negative differential between giving to black and white testers concluded in Result 4 is robust: both the non–parametric and parametric results are robust to adjustments for multiplicity (Hypotheses 1, 7, 10, 16, 17, Table 17 and Hypotheses 1–5, Table 18). The difference in giving between South–Asian and black testers is reasonably robust, but only remains significant when all observations from both the *Short* and *Long* distance treatments are pooled (Hypothesis 9, Table 17). Second, although the non–parametric results in support of Result 5 are not found to be significant once adjusted (Hypotheses 10–12, Table 17), the parametric results are found to be robust (Hypotheses 19–21, Table 18). However, Result 6 does not appear to be as robust as Results 4 and 5 (Hypotheses 22 and 28, Table 17).

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 earnings expectations stemming from bargaining with passengers explain the drivers' behaviour?; (3) can social pressure explain the extent of giving?

To examine questions (1) and (2) we conducted a complementary survey of 50 taxi drivers from ranks used within the study, 65 passengers that were queuing for a taxi, and observed the behaviour of 97 passengers entering taxis from a rank.[22] To address (1) we asked drivers the number of daily journeys they take, how many of these journeys are from taxi ranks and what they believe the average fare is. Drivers were also asked about the expected fare of an example *Short* and *Long* distance journey, where the example journeys were journeys that we used within the study. They were asked if they would be willing to bargain over the journey specified *before* the journey began, and the lowest fare they would accept if they were willing. In addition, they were asked if they would be willing to bargain with a passenger who was inside the taxi.[23] 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 14, Panel A, and the passenger responses and the observation results are presented in Panel B.

The responses in Table 14 highlight two main points relating to (1). First, the vast

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

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

Panel A:	Driver Survey, $N = 50$	
	No. daily journeys	12
		(4.4)
	No. journeys that start at a rank	11
		(4.47)
	Average fare (£)	6.41
		(1.4)
	Modal Taxi Model	LTI TXII
Short distance journeys	Expected fare (£)	6.17
		(0.778)
	Willing to bargain? (1 if yes)	0.06
		(0.242)
	Lowest fare if willing (£)	4.73
		(2.11)
	Willing to bargain inside the taxi? (1 if yes)	0.04
		(0.2)
	Upon completion	Return to home rank (73%)
		Return to a diff. rank$^{\diamond}$ (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 14: Driver and Passenger Survey Responses

majority of taxi journeys are taken from ranks (92%). This suggests that giving to the passenger, by continuing to drive away from the rank, is not done at the drivers' convenience. On the contrary, driving away from the rank is the same as driving away from the next passenger, and therefore is costly. Second, only 10% 'Cruise' upon completing a journey, with the vast majority returning to a home rank and only a minority returning to a different rank: drivers reported returning to busy city

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

In relation to (2), from Table 14 note that only 6% of drivers said they would bargain with a passenger before the passenger was inside the vehicle for the *Short* distance journey, and only 12% in the *Long* distance journey; the lowest fare they would accept is also above the amount our testers could afford. The majority would refuse to bargain with them prior to the journey beginning, and only 2 reported they would bargain with a passenger mid–journey. Their expected fare estimates are also in–line with our own calculations. Our survey and observation of passengers also shows the desire to negotiate is limited, with only a single passenger observed attempting to bargain with a driver and only 2 reporting that they did bargain with drivers over fares. Therefore, it seems unlikely that driver giving is the result of earnings expectations stemming from passenger bargaining, as the vast majority of journeys are not bargained over.

Question (3) implies that drivers are concerned about appearing unkind to the passenger, and give despite having a preference not to. This would resonate with the conclusion of DellaVigna *et al.* (2012). However, *not* giving away goods and services for free in a market setting is unlikely to be perceived as unkind. This contrasts with charitable giving, where giving to those who need it might be viewed as a normative action. Further, in the context of our study, passengers could easily have taken an alternative and cheaper mode of transport, or could have walked the final portion of the journey they couldn't afford.[24]

8 Conclusion

We report evidence that the majority of taxi drivers express other–regarding preferences in a competitive market setting, and find little evidence of the reputational concerns that are often used to explain such behaviour. Our conclusions contrast with the results of previous prominent field experiments and standard economic theory, but resonate with the results of numerous laboratory experiments and behavioural theories of social preferences. Within a highly competitive market setting, we observe individuals behaving altruistically.

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

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

when faced with a black passenger. Weaker evidence that South–Asian drivers' preferences are reduced when faced with a white passenger are also reported.

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

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

References

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

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

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

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

Bolton, G. E. & Ockenfels, A. (2000), 'ERC: A Theory of Equity, Reciprocity, and Competition', *American Economic Review* **90**(1), 166–193.

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

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.

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

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.

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.

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

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

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

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

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

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

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

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

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

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* pp. n/a–n/a.

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

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

Holm, 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.

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

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.

Riach, P. A. & Rich, J. (2002), 'Field experiments of discrimination in the market place', *The Economic Journal* **112**(483), F480–F518.

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

Schmidt, K. M. (2011), 'Social preferences and competition', *Journal of Money, credit and Banking* **43**(s1), 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.

Whitt, S. & Wilson, R. K. (2007), 'The dictator game, fairness and ethnicity in postwar Bosnia', *American Journal of Political Science* **51.3**, 655–668.

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

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

UNIVERSITY OF EXETER

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

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

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

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

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

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

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

Email:

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

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

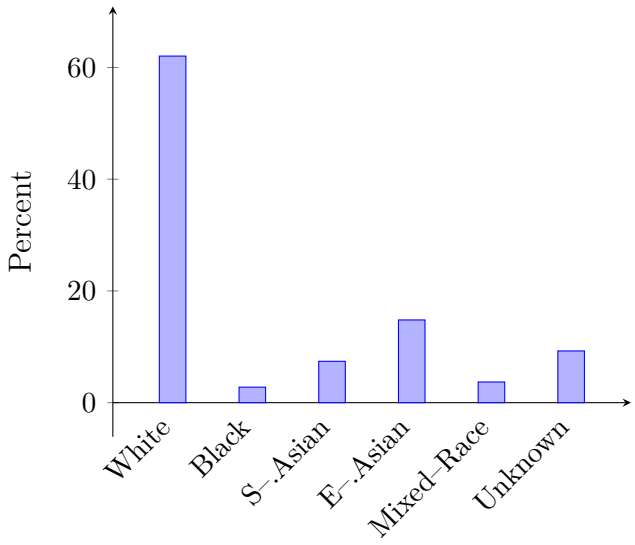
$$\Pr[x^* = \bar{x} | \alpha, \theta, a, b, c, d, e, \sigma] = \int_{\epsilon_{\bar{x}-1}}^{\infty} f(z)dz = 1 - F(\epsilon_{\bar{x}-1}). \tag{6}$$

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

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

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

B.2 Raters' ethnic demographics



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

Figure 5: Distribution of the Raters' Self–Reported Ethnicity

B.3 Testers' appearance characteristics

Panel A: *Appearance Correlations, Pooled*

	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.2839*	1.0000			
Friendly	-0.6507*	0.4358*	1.0000		
Trustworthy	-0.6641*	0.4319*	0.7835*	1.0000	
Wealthy	-0.3067*	0.4724*	0.3612*	0.3948*	1.0000

Note: 660 observations.

Panel B: *Appearance Correlations, White Testers*

	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.1501*	1.0000			
Friendly	-0.6051*	0.3766*	1.0000		
Trustworthy	-0.6189*	0.3383*	0.7746*	1.0000	
Wealthy	-0.2077*	0.5125*	0.2925*	0.3567*	1.0000

Note: 360 observations.

Panel C: *Appearance Correlations, Black Testers*

	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.4602*	1.0000			
Friendly	-0.6662*	0.5530*	1.0000		
Trustworthy	-0.6497*	0.6203*	0.7554*	1.0000	
Wealthy	-0.3814*	0.3674*	0.4039*	0.4190*	1.0000

Note: 210 observations.

Panel D: *Appearance Correlations, S.Asian Testers*

	Aggressive	Attractive	Friendly	Trustworthy	Wealthy
Aggressive	1.0000				
Attractive	-0.2330*	1.0000			
Friendly	-0.5211*	0.2552*	1.0000		
Trustworthy	-0.5945*	0.2871*	0.6151*	1.0000	
Wealthy	-0.0586	0.2681*	0.0538	0.1408	1.0000

Note: 90 observations.

Note: * indicates significance at the 5% level.

Table 15: Tester Appearance Correlations

B.4 Reduced form ethnic interactions

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

Note: Standard errors in parentheses. ***, ** and * indicate significance at the 1%, 5% and 10% level. The estimates are obtained using observations from all treatments, but we exclude observations where the driver is black. The number of observations fall slightly as more controls are included due to missing entries. Appearance Controls include measures of the Testers' aggressiveness, attractiveness, friendliness, trustworthiness and wealthiness, as outlined in Section 4.1. Observations with a white driver and a white Tester are taken as the baseline.

Table 16: The Effects of Ethnic Interactions on Giving

B.5 Robustness checks

#	*Alt. Hypothesis*	*Family*	*Outcome*	*Unadjusted*	*Adjusted*
Result 4					
1	H_A: White ≠ Black	*No–Reputation, Short*	Giving, £	0.001***	0.003***
2	H_A: White ≠ S.–Asian			0.37	0.74
3	H_A: S.–Asian ≠ Black			0.06*	0.12
4	H_A: White ≠ Black	*No–Reputation, Long*	Giving, £	0.396	1.00
5	H_A: White ≠ S.–Asian			0.88	0.88
6	H_A: S.–Asian ≠ Black			0.47	0.94
7	H_A: White ≠ Black	*No–Reputation, pooled*	Giving, %	0.005***	0.015**
8	H_A: White ≠ S.–Asian			0.311	0.0311
9	H_A: S.–Asian ≠ Black			0.025**	0.05**
10	H_A: White ≠ Black	*Reputation, Short*	Giving, £	0.0003***	0.0009***
11	H_A: White ≠ S.–Asian			0.13	0.26
12	H_A: S.–Asian ≠ Black			0.622	0.622
13	H_A: White ≠ Black	*Reputation, Long*	Giving, £	0.566	1.00
14	H_A: White ≠ S.–Asian			0.46	0.46
15	H_A: S.–Asian ≠ Black			0.45	0.90
16	H_A: White ≠ Black	*Reputation, pooled*	Giving, %	0.0005***	0.0015***
17	H_A: White ≠ S.–Asian			0.003***	0.006***
18	H_A: S.–Asian ≠ Black			0.90	0.90
Result 5					
19	H_A: White ≠ Black	*No–Reputation, pooled*	Journey Completion	0.045**	0.135
20	H_A: White ≠ S.–Asian			0.793	0.793
21	H_A: S.–Asian ≠ Black			0.088*	0.176
Result 6					
22	H_A: *No Rep.* ≠ *Rep.*, White	*No–Rep. vs Rep., Short*	Giving, £	0.055*	0.165
23	H_A: *No Rep.* ≠ *Rep.*, Black			0.593	0.593
24	H_A: *No Rep.* ≠ *Rep.*, S.–Asian			0.278	0.556
25	H_A: *No Rep.* ≠ *Rep.*, White	*No–Rep. vs Rep., Long*	Giving, £	0.61	1.00
26	H_A: *No Rep.* ≠ *Rep.*, Black			0.624	0.624
27	H_A: *No Rep.* ≠ *Rep.*, S.–Asian			0.5 06	1.00
28	H_A: *No Rep.* ≠ *Rep.*, White	*No–Rep. vs Rep., pooled*	Giving, %	0.053*	0.159
29	H_A: *No Rep.* ≠ *Rep.*, Black			0.397	0.397
30	H_A: *No Rep.* ≠ *Rep.*, S.–Asian			0.15	0.3

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

Table 17: Adjusted p-values – Non–Parametric Testing

Table	#	*Result*	*Model*	*Explanatory Variable of Interest*			
				Black	*South–Asian*	*Male*	m
	1		(1)	0.002***	0.615		2
	2		(2)	0.003***	0.279	0.106	3
Table 10	3	Result 4	(3)	0.000***	0.129	0.06*	3
	4		(4)	0.005***	0.287	0.126	3
	5		(5)	0.001***	0.981	0.462	8
	6		(1)	0.022**	0.197		2
	7		(2)	0.048**	0.153	0.246	3
Table 11	8	Result 5	(3)	0.049**	0.312	0.286	3
	9		(4)	0.057*	0.242	0.188	3
	10		(5)	0.056*	1.00	1.00	8

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

Table 18: Adjusted p–values – Parametric Testing