Mohammed Goes House-Hunting: An Experimental Study of Discrimination

Itzhak Rasooly*

Supervisor: Professor Johannes Abeler

Submitted in partial fulfilment of the requirements for the degree of

Master of Philosophy in Economics

Department of Economics

University of Oxford

Trinity Term 2018

* I would like to thank my supervisor, Johannes Abeler, for his support and guidance throughout the completion of this thesis. I would also like to thank Jasmine Theilgaard for extremely useful feedback. Nat Levine, Adam Brzezinski and Jesper Åkesson provided valuable comments. Word count: 16,520.

Abstract

We conduct a field experiment in order to investigate discrimination against Muslims in the UK rental market. We find that Muslims are significantly less likely to receive a viewing, an extremely robust result that persists across genders, cities and landlord and property types. Further, there is evidence that even those landlords who do not formally discriminate are nonetheless less polite to Muslim applicants. Neither of the leading theories appear well suited to explain the discrimination observed in this study. On the one hand, discrimination completely disappears once we restrict attention to Muslim landlords, a fact that is difficult to reconcile with statistical models of discrimination. On the other hand, discrimination does not increase with interaction, in apparent contradiction of Becker's 'taste based' model. We argue that a re-interpretation of the Becker model can explain our findings.

Contents

1 Introduction

Experimental studies have documented the existence of discrimination against Muslims in a range of contexts. However, they have typically proven less successful in explaining why this discrimination occurs. Many studies do not even attempt to link their findings with any of the leading theories of discrimination (Bertrand & Duflo, 2017). Furthermore, as we argue in Section 3, many of the existing approaches to identifying the source of discrimination are theoretically questionable and rest on strong yet undefended auxiliary assumptions.

One obstacle to understanding why discrimination occurs is a lack of data on the personal characteristics of potential discriminators. When sending fictitious applications to employers, estate agents or other groups, experimenters typically know very little about the identity of the individual assessing the application. In this study, we circumvent this difficulty by conducting a field experiment on the flat sharing website SpareRoom. After sending 1,368 fictitious applications in a 'matched design' that pairs each landlord with a non-Muslim and Muslim applicant, we then record the property type, landlord ethnicity and whether the landlord will be living with the applicant – three variables that allow us to test various theories of discrimination.

The results demonstrate the existence of widespread discrimination against Muslims. Non-Muslims are between 30% and 32% more likely to be offered a viewing, depending on how one codes the responses. This result is extremely robust and remarkably persistent. We observe discrimination by both male and female landlords against both male and female Muslim applicants. Discrimination spans a range of landlord and property types. Although some cities receive as few as 126 applications, we find statistically significant discrimination in every city in the sample.

We also find evidence that even those landlords who appear to treat the Muslim and non-Muslim similarly (by accepting or rejecting both) are nonetheless less polite to the Muslim. However, this result turns out to be sensitive to the choice of weights in our politeness index and loses significance once we take account of the multiplicity of our hypothesis tests. Hence, this finding should be viewed as merely suggestive, and as a prompt for future research.

In order to understand why the discrimination occurs, we begin by studying the European (but non-English) names in the sample. If the discrimination were 'just' xenophobia, then it would disappear in cases where the Muslim name is paired with European names. However, in such cases discrimination rates remain as high if not higher than in the rest of the sample. Similarly, by examining the case when a Muslim name is paired with lower class names we can reject the hypothesis that discrimination occurs because Muslims are perceived as lower class.

According to models of 'statistical discrimination', discrimination occurs not because discriminators are prejudiced but rather because they rationally use ethnicity to infer information about important yet imperfectly observed applicant attributes. To test this, we examine whether discrimination varies with the ethnicity of the landlord, finding that it completely disappears once we restrict attention to landlords with Muslim names. As we argue, this finding is hard to reconcile with even permissive models of statistical discrimination which allow for variation in the information to which landlords have access.

We also test Becker's (1957) hypothesis that discrimination results from a dislike of interaction with minority groups. First, we compare live-out landlords and estate agents with those decision-makers who will have to live with (and

therefore substantially interact with) the applicant. Surprisingly, we find that the latter group actually discriminate less, although the difference is not statistically significant. We also compare properties that are offered as a whole with those that are offered as a house or flat share. Again, we find that more interaction actually means less discrimination, in apparent contradiction of Becker's classic theory.

Although we cannot definitively identify the source of the discrimination we uncover, we can suggest some theories that are consistent with our data. One option is that Muslim landlords have (relatively) more favourable beliefs about Muslim tenants, not because of the information to which they have access but because they are (relatively) biased in favour of their own group. Another option is that landlords have a desire to help those who belong to their own group (or harm those who do not), whether they interact with them or not. Therefore, while we find no evidence for Becker's interaction hypothesis, we do propose explanations that are close in spirit to Becker's notion of 'taste based' discrimination.

The rest of the study is organised as follows. Sections 2 and 3 review the theoretical and empirical literature on discrimination. Section 4 presents the experimental design. Sections 5 and 6 contain the findings and Section 7 discusses what might explain them. Section 8 presents extensions and Section 9 concludes with a discussion of the study's limitations and directions for future research.

2 Theories of Discrimination

2.1 Defining Discrimination

By 'discrimination', we mean any differential treatment that arises by virtue of membership of a certain group. In other words, we define discrimination as the

causal effect of group membership on some outcome of interest. While simple, a few comments on this definition are in order.

First, while this definition captures many phenomena which are widely considered to be objectionable, it also captures phenomena which few would oppose. If a film executive only considers female actresses when auditioning for a female role, or a person only considers male candidates when looking for a husband, then they are discriminating (on our definition) – and yet few would criticise them for doing so. In this sense, our definition is not 'morally loaded'.

Second, the definition here is neutral as to the source of discrimination, allowing it to be motivated by animus, statistical considerations or anything else. In this respect, we differ from other superficially similar definitions in the literature. For example, Heckman (1998) writes that 'racial discrimination is said to arise if an otherwise identical person is treated differently by virtue of that person's race or gender, and race or gender by themselves have no direct effect on productivity'. By this, he means that if an employer refuses to hire a certain group because they tend to be unproductive, that is not discrimination – hence the difficulty of 'identifying discrimination' from experimental studies.[1] In contrast, our definition would consider this discriminatory.

---

[1] In fact, Heckman is extremely misleading about what constitutes discrimination. Just after the quotation given above, he writes that discrimination is 'a causal effect defined by a hypothetical *ceteris paribus* conceptual experiment–varying race but holding all else constant' (p. 102). From the text, it is clear that one of the things we need to hold constant is employer beliefs about productivity which (Heckman assumes) may well depend on the employee's race. This is, after all, the basis of one of his main criticisms of audit studies: that they fail to hold beliefs about unobservable quality constant when they experimentally manipulate gender or ethnicity. But this is not what is meant by a causal effect. When considering the causal effect of weather on ice cream consumption, we do not hold constant the number of people who go to the beach. More generally, if X determines Z via an intermediate factor Y, we do not typically hold Y constant when considering the causal effect of X on Z. Thus, Heckman's definition of the causal effect of

2.2 Explaining Discrimination

Having defined discrimination, we can then ask what explains it. One natural explanation is that decision-makers dislike interacting with members of certain groups (Becker, 1957).[2] This leads them to act as if they are willing to pay a price to avoid such interactions. Since discrimination here is generated by preferences (or 'tastes'), this has been dubbed 'taste-based discrimination'.

However, discrimination need not have anything to do with discriminatory preferences. According to statistical models of discrimination (Phelps, 1972; Arrow, 1973; Aigner & Cain, 1977) employers discriminate not because they are prejudiced but because they use group membership as a signal of other important but unobserved characteristics. In other words, discrimination arises not from discriminatory preferences but as the optimal solution to a prediction problem under incomplete information.

To see how this could work in the context of the rental market, suppose that a tenant's 'quality' (whether they pay their bills on time, keep the property in good condition, etc.) can be summarised by a number, Q. Suppose also that Q is normally distributed with a religion-specific mean μ_x and variance σ_x^2, indexed by $x = m, n$ depending on whether the tenant is Muslim (m) or non-Muslim (n). Finally, suppose that while landlords cannot observe a tenant's quality, they can observe the quality of their message M = Q + u where u is independent of Q and distributed normally with mean 0 and variance σ_u^2. That is, although landlords

race as the effect on outcomes holding employer beliefs constant is rather unusual, to say the least.

[2] Throughout, we will use the term 'decision-makers' to refer to potential discriminators, whether they be employers, landlords or some other group.

cannot observe a tenant's quality, they do have access to a noisy but unbiased estimator of that quality. (The set-up here is based on Aigner & Cain, 1977.)

Given these assumptions, M is distributed normally with mean μ_x and variance $\sigma_x^2 + \sigma_u^2$. Furthermore, it is straightforward to compute Cov(M, Q) = Cov(Q, Q) = σ_x^2. To summarise, then, messages and quality are jointly distributed as follows:

$$\begin{pmatrix} \mathrm{M} \\ \mathrm{Q} \end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_x \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_x^2 + \sigma_u^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 \end{bmatrix} \right)$$

Let $\gamma_{\mathrm{x}} \stackrel{\text{def}}{=} \frac{\sigma_{\mathrm{x}}^2}{\sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{u}}^2}$. Using a well-known fact about the conditional distribution of a multivariate normal, we obtain:

$$E[\mathrm{Q} \mid \mathrm{M} = m] = (1 - \gamma_x)\mu_x + \gamma_x m \tag{1}$$

This says that expected quality given the message (which we assume equals landlord expectations) is a weighted average of the quality of the message and the mean quality of the applicant's group.[3] As one would expect, as messages become totally uninformative, i.e. $\sigma_u^2 \longrightarrow \infty$, landlord expectations simply tend to group means. Equally unsurprisingly, if messages are perfect predictors of quality, i.e. $\sigma_u^2 = 0$, then the expected quality of a tenant bears no relation to the average quality of their group. This highlights the need for imperfect information in models of statistical discrimination.

[3] Of course, γ_x is also the limit of the OLS slope estimator with one explanatory variable and an intercept. Hence, even if landlords do not believe that these variables are normally distributed, they will act approximately as if they do if they use OLS estimation on a sufficiently large sample.

To see the first and most obvious point, suppose that the variance in quality is the same for both groups. In that case, the only way that landlords can distinguish two 'identical' applicants (who send the same quality message but have different religions) is through the average quality of their group. If one applicant belongs to a group with higher mean quality, that applicant must be (weakly) favoured by the landlord.

Suppose next that the groups have the same average quality but different variances. In that case, landlords will use different weights on the messages from different groups. If two applicants both send (identical) above average messages, the landlord will favour the applicant for whom messages better predict quality. Since γ_x is increasing in σ_x^2, this is the applicant who belongs to the religion whose quality varies the most. On the other hand, if the messages are below average, the landlord will (weakly) prefer the applicant whose quality varies the least. If messages set equal to the average quality, as we attempted to do in the experiment, then this effect disappears.

Allowing for landlord risk aversion provides still another way for the covariance terms to come into play. To see this, assume that landlords have exponential utility, which means that their expected utility is given by:

$$\mathrm{E[U(Q)]} \;=\; \mathrm{E[Q]} \;-\; \frac{r}{2}\mathrm{Var[Q]}$$

Accordingly, the 'riskiness' of a tenant is proportional to the variance of their quality. (By the Arrow-Pratt approximation, this is roughly true even without exponential utility.) Using another well-known result from normal distribution theory, we obtain:

$$\mathrm{Var[Q} \mid M = m] \;=\; \frac{\sigma_x^2\;\sigma_u^2}{\sigma_x^2 + \sigma_u^2} \tag{2}$$

The first thing to notice is that the conditional variance is increasing in σ_x^2 since:

$$\frac{\partial \mathrm{Var}(\mathrm{Q} \mid M = m)}{\partial \sigma_x^2} = \frac{\sigma_u^4}{(\sigma_x^2 + \sigma_u^2)^2} > 0$$

By symmetry, it must also be increasing in σ_u^2. The implication is straightforward: if an applicant comes from a group whose quality varies greatly or is measured with higher variance error, then that applicant's quality will itself be high variance. Since landlords dislike variance, this makes it (weakly) less likely that the applicant will be chosen.

In summary, then, the model suggests a number of channels through which discrimination might occur:

1. Discrimination tends to occur against the group with the lower average quality.

2. If messages are better (worse) than average, discrimination tends to occur against the group whose quality is worse (better) predicted by messages.

3. If landlords are risk-averse, discrimination tends to occur against the group whose quality is more variable or measured with greater error.

One might wonder about the case in which messages fail to communicate anything except an applicant's group status. That case is even more straightforward. Without risk aversion, statistical discrimination can only emerge through differences in group means. With risk aversion, landlords might also statistically discriminate against the group whose quality is more variable.

At this stage, a key feature of this model should be highlighted. In the model, we assumed that landlords correctly know the joint distributions of quality and are able to correctly deduce the distribution of quality conditional on the quality of a message. In practice, however, either of these assumptions could be relaxed. Landlords might have erroneous beliefs about the distributions of quality, or they might fail to condition correctly on the available information, or both. Following England (1992), we will call discrimination resulting from such processes 'error discrimination' and reserve the term statistical discrimination for models in which agents are rational and fully attentive to the available information.[4]

It should also be stressed that taste-based discrimination, statistical discrimination and error discrimination are not the only options. To provide just one example, agents might discriminate not because they dislike interacting with a minority group but because they enjoy helping their own group ('ethnic homophily'). While this explanation could be formalised with the algebra of the Becker model (Becker, 1957), it has nothing to do with interaction *per se*. We return to this point in Section 7.5.

3 Empirical Studies of Discrimination

This study contributes to two overlapping literatures: the literature on whether discrimination exists and the literature on why it occurs. We discuss each of these in turn.

[4] Unfortunately, the literature is not consistent on whether it terms such models statistical discrimination. On the one hand, Aigner and Cain (1977) assume that employer estimates are at least unbiased, an assumption they justify with reference to the "forces of competition" (p.177). On the other hand, Arrow (1973) explicitly discusses the possibility of statistical discrimination based on inaccurate beliefs, a definition followed by some recent papers (such as Zussman, 2013).

3.1 Does Discrimination Exist?

While some evidence of discrimination comes from observational data (Altonji & Blank, 1999), field experiments provide researchers with a cleaner way to identify the causal effect of group membership. Although field experiments on discrimination have existed at least since Joel and Prescott Clark (1970), they have exploded in popularity since Bertrand and Mullainathan's (2004) study of discrimination against African-Americans in the US labour market. In recent years, they have uncovered discrimination on the basis of class, gender, race, religion, caste, sexuality, attractiveness, criminal record, unemployment duration, age and disability, to provide a non-exhaustive list. Bertrand and Duflo (2017) provide a useful overview.

Field experiments of discrimination come in many forms. One distinction is between matched and unmatched designs. In matched studies, such as Ayres and Siegelman (1995), the experimenter sends two applications to each decision-maker, one of which is from the minority group and one of which is not. In unmatched studies (such as Edelman, Luca & Svirsky, 2017), the experimenter randomises group membership when making applications, relying on large samples to ensure that all groups apply to roughly the same decision-makers. Another useful distinction is between audit and correspondence studies. In audit studies (such as Hebl, Foster, Mannix & Dovidio, 2002), applications are made by trained actors. In practice, such audit studies typically use matched designs, although this need not be the case. Meanwhile, correspondence studies (which comprise most of the current literature) use paper or online applications. The costs and benefits of these various experimental designs will be discussed in Section 4.

Although attention in the US has focused on discrimination against African-Americans, there has also been considerable focus on discrimination against Muslims in Europe. Looking at labour markets, researchers have found

discrimination against Muslim applicants in the Netherlands (Blommaert, Coenders & van Tubergen, 2013), Germany (Kaas and Manger, 2012; Bartoš, Bauer, Chytilová & Matějka, 2016), Belgium (Baert, Cockx, Gheyle & Vandamme, 2013) and the US (Wright, Wallace, Bailey & Hyde, 2013; Acquisti & Fong, 2015). In these studies, non-Muslim applicants are between 6% and 62% more likely to receive a call-back. With respect to rental markets – the focus of this study – researchers have found anti-Muslim discrimination in Sweden (Ahmed & Hammarstedt, 2008; Ahmed, Andersson & Hammarstedt, 2010; Bengtsson, Iverman & Hinnerich, 2012; Carlsson & Eriksson, 2014), the US (Carpusor & Loges, 2006), Spain (Bosch, Carnero & Farré, 2010) and Italy (Baldini & Federici, 2011). However, researchers have not replicated such results across cities in the UK, leaving a gap in the literature which this study aims to fill.

More importantly, the studies which already exist are typically unable to examine how discrimination depends on the personal characteristics of the decision-maker.[5] This is unsurprising: when applying to a firm for a job or to an estate agent for an apartment, one rarely learns very much about the particular person making the decision about one's application. In contrast, the present study allows us to examine the effect of two variables: the *ethnicity* of the landlord and the amount of *interaction* the landlord will have with the applicant (which hinges on whether they would live with the applicant or merely show the around the property). As will later be argued, this allows us to test theories of taste-based and statistical discrimination.

[5] Two partial exceptions are Ahmed & Hammarstedt (2008) and Carlsson & Eriksson (2014) who examine whether landlords with an ethnic minority name are less likely to discriminate (obtaining conflicting results). However, neither study addresses the question of whether *Muslim* landlords are less likely to discriminate – as opposed to landlords with minority names – and therefore neither study explicitly tests for ethnic homophily, as we do here. Moreover, this is to our knowledge the first study to investigate the role of landlord/applicant interaction.

3.2 What Explains Discrimination?

Although the evidence just discussed points overwhelmingly to the existence of discrimination against Muslims and other groups, it is far less clear why this occurs. While a number of different empirical strategies have been pursued to throw light on this question, many of them suffer from quite serious (but little acknowledged) faults. Identifying the flaws in some of these methods motivates the alternative strategies pursued in this study.

3.2.1. Varying Information

One popular strategy is to examine how discrimination varies with the amount of information provided by the experimenter (Ahmed et al., 2010; Kaas & Manger, 2012). The idea is that if discrimination is statistical, it should decrease when more information is given, providing a test of the theory. As Ahmed et al. (2010) put it:

> If we still observe differences between Fredrik and Mustafa [after providing additional information], this would be an indication of that lack of information about employment, education, and marital status is not the reason for statistical discrimination, and that instead it is more likely that landlords are engaging in preference-based discrimination. (p. 85)

Is this correct? Certainly, statistical discrimination can take place only if there is imperfect information, which is no doubt the intuition behind such tests. However, intuitions should be checked with models, and it is regrettable that the authors of these studies fail to specify the exact model that they have in mind. Fortunately, we can investigate the issue using the benchmark model already outlined in Section 2.2.

To begin, suppose again that γ is the same for both groups and consider the difference in expected quality between a Muslim and non-Muslim applicant conditional on the message. From (1), this is given by:

$$\mathrm{E}[\mathrm{Q}_n \mid \mathrm{M} = m] - \mathrm{E}[\mathrm{Q}_m \mid \mathrm{M} = m] = (\mu_n - \mu_m)(1 - \gamma) \tag{3}$$

By providing more information, the experimenters are effectively reducing σ_u^2. In other words, they are improving the decision-maker's ability to predict tenant quality from the message. As previously noted, γ is increasing in σ_u^2, so the absolute value of (3) is (monotonically) decreasing in σ_u^2. So we see that (at least in this case) the intuition is exactly right: not only does providing perfect information ($\gamma = 1$) eliminate the gap, but providing more information always makes the gap smaller.[6]

Consider now the difference in conditional variances between the Muslim and non-Muslim applicant. From (2), this is given by:

$$\mathrm{Var}[\mathrm{Q}_n \mid m] - \mathrm{Var}[\mathrm{Q}_m \mid m] \;=\; \frac{\sigma_n^2\,\sigma_u^2}{\sigma_n^2 + \sigma_u^2} \;-\; \frac{\sigma_m^2\,\sigma_u^2}{\sigma_m^2 + \sigma_u^2} \tag{4}$$

[6] Technically speaking, this is not quite right since the amount of discrimination (in the model) does not depend on the *magnitude* of the differences in expected quality. However, this can be easily fixed by recognising that landlords evaluate expected quality with error. For example, suppose that they compare $\mathrm{E}[\mathrm{U}(\mathrm{Q}_n)]$ with $\mathrm{E}[\mathrm{U}(\mathrm{Q}_m)] + \epsilon$ where ϵ is standard normal and independent of Q. Then the probability a landlord rejects the Muslim but accepts the non-Muslim is $\Phi(\mathrm{E}[\mathrm{U}(\mathrm{Q}_n) - \mathrm{E}[\mathrm{U}(\mathrm{Q}_m))$ and thus strictly increasing in the difference in expected quality. Heterogenous landlord thresholds would provide another route to this result.

To see that (4) is monotone in σ_u^2, we differentiate it with respect to σ_u^2, obtaining:

$$\frac{\sigma_u^4}{(\sigma_n^2+\sigma_u^2)^2} - \frac{\sigma_u^4}{(\sigma_m^2+\sigma_u^2)^2} \tag{5}$$

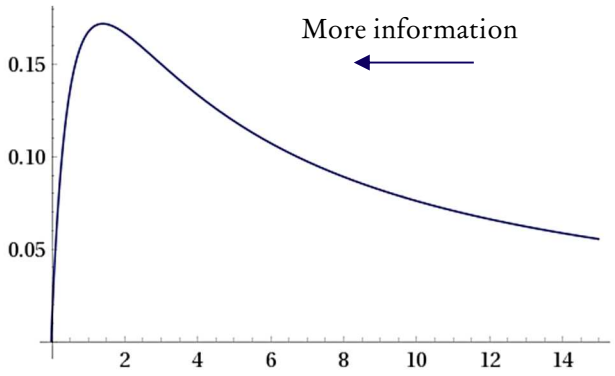
Thus, if $\sigma_n^2 > \sigma_m^2$, the derivative is negative for any σ_u^2; and if $\sigma_n^2 \leq \sigma_m^2$ the derivative is non-negative for any σ_u^2. Therefore, varying σ_u^2 cannot change the sign of (5), establishing that (4) is monotone. Since (4) is zero when $\sigma_u^2 = 0$, the absolute value of (4) must be weakly increasing in σ_u^2. Again, the news is positive: not only does providing perfect information (σ_u^2 = 0) eliminate the variance gap, but providing better information always reduces the gap.

However, a difficulty emerges once we consider the effect of unequal variances on the conditional means. To see this, suppose now that Muslims and non-Muslims have the same mean quality μ. From (1), the gap between the conditional mean of two applicants is then given by:

$$\mathrm{E}[\mathrm{Q}_n \mid \mathrm{M} = m] - \mathrm{E}[\mathrm{Q}_m \mid \mathrm{M} = m] = (\gamma_n - \gamma_m)(m - \mu)$$

In fact, this function is not monotone in σ_u^2. For example, suppose that $\sigma_n^2 > \sigma_m^2$ and $m > \mu$. Then it can be shown that the function is increasing in σ_u^2 when $\sigma_u^2 < \sigma_m$ but decreasing when $\sigma_u^2 > \sigma_m$. We plot the case for $\sigma_n^2 = 2, \sigma_m^2 = 1$ and $(m - \mu) = 1$.

Figure 1. The effect of information on the variance gap.



Since we have placed no restriction on the size of $(m - \mu)$, this effect could in principle dominate the others. If so, providing more information will initially make discrimination worse, even though it will eventually make discrimination better. While this is a troubling result, we can at least say that the effect will be small if experimenters manage to send messages that are close to the average message quality.

Unfortunately, further problems for the test arise as soon as one moves away from the world of normal distributions. In fact, it turns out to be all too easy to construct examples where providing more information makes statistical discrimination worse – even if landlords only statistically discriminate using information about group means. To construct the simplest example, suppose that there are two types (Muslims and non-Muslims) and that half of each type lives in London. In London, all Muslims have a quality of 0 and all non-Muslims have a quality of 1. Outside of London, all Muslims have a quality of 1 and all non-Muslims have a quality of 0. It follows that unconditional means (and higher moments) are the same for Muslims and non-Muslims.

In the first part of a correspondence study, experimenters send some landlords messages which reveal nothing but the applicant's religion. Since the expected

qualities of Muslim and non-Muslim applicants are the same (they both equal 0.5), landlords have no reason to statistically discriminate. In the second part of the study, experimenters send landlords messages which also reveal that the applicant lives in London. In this subset, there is a considerable disparity: all Muslims have quality 0 and all non-Muslims have quality 1. Hence, if a landlord can only accept one of the two applicants, they 'should' accept the non-Muslim; so providing more information has given landlords more reason to statistically discriminate! While deliberately stylised, this example casts considerable doubt on the rationale of the test.

3.2.2. Varying Applicant Quality

Another strategy is to examine how discrimination varies with the 'quality' of applications. This strategy has become popular since Bertrand & Mullainathan's (2004) finding that discrimination is worse against individuals with high-quality CVs (something they term the 'credentials effect'). Supposedly, this is evidence against models of statistical discrimination. As they put it:

> [Statistical discrimination] models struggle to explain the credentials effect as well. Indeed, the added credentials should lead to a larger update for African-Americans and hence greater return to skills for that group (p. 1010)

Although this strategy has been pursued many times elsewhere – for example, Nunley, Pew, Romero & Seals (2015) test how discrimination varies with the quality of an applicant's major – it is questionable. To be blunt, statistical discrimination models do not (without strong auxiliary assumptions) imply anything about the relationship between discrimination and applicant quality. Indeed, one suspects that this entire approach rests on a linguistic confusion between a high-quality signal (a signal which well predicts quality) and a signal of

high quality (a signal that suggests that the applicant is desirable on some dimension). While there is some reason to think that discrimination rates depend on the former (although see the previous section), their relation to the latter is far less clear.

To see that increasing application quality does not *imply* a reduction in discrimination (in a statistical discrimination model), consider a variant on the binary example just discussed. Suppose now that if an applicant lacks a degree (which we assume means that they are of 'lower quality'), they are of the same quality regardless of whether they are Muslim or non-Muslim. However, if an applicant has a degree, then they are of a lower quality if they are Muslim. In that case, landlords have no reason to statistically discriminate when facing applicants without degrees but do have reason to statistically discriminate (against Muslims) when facing applicants with degrees. In other words, as quality increases, statistical discrimination should increase, *pace* Bertrand and Mullainathan (2004).

Moreover, higher quality need not mean less discrimination even if variables are normally distributed (as in standard models of statistical discrimination). To see this, we extend the model already developed to include two quality types: high and low. Doing so is straightforward: in addition to indexing means and variances by religion, we now also index them by quality. In terms of notation, let Q_y^x denote the quality of an applicant with religion x and quality y, where $x = m, n$ as before and $y = h, l$ for high and low-quality applicants respectively. To focus on the most important issue, assume that landlords are risk neutral (so only care about expected quality) and that variances are the same for Muslim and non-Muslim applicants (although may be different for high and low-quality applicants).

From (1), discrimination against high-quality applicants turns on:

$$\mathrm{E}[\mathrm{Q}_n^h \mid \mathrm{M} = m] - \mathrm{E}[\mathrm{Q}_m^h \mid \mathrm{M} = m] = (\mu_n^h - \mu_m^h)(1 - \gamma^h)$$

Similarly, discrimination against low-quality applicants turns on:

$$\mathrm{E}[\mathrm{Q}_n^l \mid \mathrm{M} = m] - \mathrm{E}[\mathrm{Q}_m^l \mid \mathrm{M} = m] = (\mu_n^l - \mu_m^l)(1 - \gamma^l)$$

The difference in discrimination rates (between high and low-quality applicants) is therefore determined by:

$$(\mu_n^h - \mu_m^h)(1 - \gamma^h) - (\mu_n^l - \mu_m^l)(1 - \gamma^l) \tag{6}$$

The first thing to notice is that (6) does not depend on quality. Suppose that the quality of all high types were to increase by some constant c. Then μ_n^h and μ_m^h would each increase by c, leaving the difference unchanged. Furthermore, γ^h would not change since it depends only on covariances, which are invariant to adding constants. Plainly, then, the whole issue of quality is a red herring.

While rather obvious, this point does not seem to have been appreciated by the authors of these tests in the literature. For example, Nunley et al. (2015) spend considerable time arguing that business majors are 'higher quality' than non-business majors, which supposedly implies that discrimination should be lower amongst business majors. However, as (6) makes clear, the relevant objects are: 1) the difference in 'signal quality' between business and non-business majors (which depends on the variances of quality and measurement error for these two types) 2) the difference in the 'mean ethnic quality gap' between high and low skilled applicants. Since Nunley et al. do not discuss any of these objects, it seems that they are unaware of the assumptions on which their test relies.

Of course, it might turn out that (6) is negative, as they require. For example, if the mean differences are the same for high and low types and the quality of business majors fluctuates more than the quality of non-business majors, then (assuming the variance of error is the same for all groups) discrimination will be lower for business majors. However, any test of statistical discrimination that relies on so many contestable auxiliary assumptions is unlikely to be compelling. At any rate, future work in this area needs to at least spell out the assumptions behind the test and provide them with some empirical justification.

3.2.3. Surveys and Interviews

Recently, some economists have begun to explore the determinants of discrimination by investigating how discrimination varies with explicit or implicit attitudes. For example, Charles & Guryan (2008) find that racial wage gaps in the United States are well predicted by a prejudice index constructed from answers to the General Social Survey. While based on observational data, this finding is robust to the inclusion of various controls. Charles and Guryan therefore conclude that "racial prejudice among whites accounts for as much as one-fourth of the gap in wages between blacks and whites" (p. 805).

Assuming that the effect they identify is causal (an issue which need not concern us here), these results are inconsistent with the standard model of statistical discrimination. In the model, all agents have the same beliefs so there is no reason for discrimination to be predicted by a prejudice index. However, one could instead imagine that agents have access to different information and therefore discriminate according to different statistical models. In that case, discrimination could be predicted by a prejudice index if that index also picks up on variation in information.

To illustrate, consider one of the questions whose answers contribute to the index:

> If you were driving through a neighbourhood in a city, would you go out of your way to avoid a black section? (p. 806)

While answers to this question will presumably pick up attitudes towards African-Americans, they could also pick up information about ethnic differences in crimes rates – which in turn might drive discrimination. In other words, it is open to the defender of statistical discrimination to claim that the 'prejudice index' is really an information index.

To be sure, the most plausible explanation of affirmative answers to questions like these may well be a dislike of African-Americans. However, that is not something that we learn from the study. Rather, it is a prior belief which we make take to the study and use to interpret the results. If we were to approach the study with a prior belief that answers reflect differences in information, we could instead view the results as evidence of statistical discrimination. Hence, the Charles and Guryan approach does not really provide us with an independent test of these theories.

Interestingly, this criticism is to some extent avoided by a recent study which investigates discrimination against Arab-Israelis (Zussman, 2013). Discrimination turns out to be especially well predicted by answers to the following question in a follow-up survey:

> The Arabs in Israel are more likely to cheat than the Jews. (p. 436)

Importantly, this statement remains a strong predictor even after controlling for answers to other questions which plausibly measure (normative) attitudes towards Arab-Israelis. This does suggest that discrimination is driven by the empirical belief that Arab-Israelis cheat, although of course it does not show that such beliefs are warranted by the evidence.

3.2.4. Summary

Correspondence studies have established the existence of discrimination against Muslims in rental and labour markets across several countries. However, they have been less successful in throwing light on why discrimination occurs; and some of the strategies commonly employed are theoretically questionable. This motivates our alternative approach: examining whether discrimination depends on the ethnicity of the landlord and the extent to which they interact with the tenant.

4 Experimental Design

Choosing how to conduct a field experiment on discrimination forces the researcher to make a large number of choices. We discuss each of these in turn:

Correspondence vs. audit study. This choice was rather straightforward since an audit study (which would involve either calling landlords or visiting estate agents in person) would require far more time than simply sending applications online. In addition, using an audit design may introduce unconscious bias (Heckman, 1998). Therefore, we followed most papers in the current literature by opting for a correspondence study.

Varying application quality. In an influential paper, Neumark (2012) shows that one can identify the proportion of discrimination that is taste-based by

varying the quality of applications. The idea is to use variation in application quality to estimate the variances of unobservable quality and thereby remove the effect of any variance-based discrimination. He therefore recommends that all future field experiments have at least two quality types, a recommendation that has since been accepted by some researchers (see for instance Carlsson, Fumarco and Rooth, 2014).

We chose not to implement the Neumark approach because it rests on the extremely strong identifying assumption that mean quality is the same across groups. In practice, it is hard to imagine a scenario in which a researcher could know that two groups have the same mean quality but not that they have the same variances. Therefore, we decided to send just one quality of application.

Matched vs. unmatched designs. The relative merits of matched and unmatched designs involve both statistical and non-statistical considerations. On the statistical side, matched designs require twice as many applications to achieve the same number of independent observations, which effectively halves the sample size. On the other hand, the positive association that one typically observes between the replies of a given landlord reduces the variance of the estimated difference in proportions, increasing statistical power (Agresti, 2007, p. 245). In order to resolve this trade-off, we calculated the samples required to give us 95% power using the matched and unmatched designs, finding that the matched design would require just 1/5 the sample size of the unmatched design.[7] This benefit more than offsets the cost of effectively halving the sample.

[7] When carrying out power calculations, we assumed that we would analyse results from the matched and unmatched designs using the McNemar test (described in Section 5) and chi-squared test of independence, respectively. The formula for the McNemar test is given by Vuolo, Uggen and Lageson (2016). We set $\alpha = 1 - \beta = .05$, $p_{10} = .15$ and $p_{01} = .05$, where p_{10} is the probability that a landlord rejects the Muslim but accepts the non-Muslim and p_{01} is the probability that she does the opposite.

There were also non-statistical considerations. Sending two similar messages to each landlord at similar times might arouse their suspicions, leading them to behave in an unrepresentative manner (Vuolo, Uggen & Lageson, 2018). However, since most landlords presumably receive a number of messages, we judged this risk to be slim. Furthermore, a matched design ensures that the Muslim and non-Muslim applicants contact exactly the same landlords to express interest in exactly the same properties, making the causal identification more compelling. In contrast, unmatched designs need to rely on randomisation and large samples (and possibly controls) to ensure that Muslims and non-Muslims apply to similar landlords. For both statistical and non-statistical reasons, then, we opted for a matched design.

This same power calculation suggested that we would need a sample of at least 250 landlords. However, we wished to be able to detect discrimination amongst the sub-set of Muslim landlords who we anticipated to form 10% of the sample. We therefore aimed for a sample of 2,500 landlords, which meant sending 5,000 messages. In the event, the experiment was interrupted mid-flow (see below) so the sample ultimately only comprised of 684 landlords (to whom we sent 1368 messages).

Platform. Since we wanted to test the importance of interaction, we needed to use a flat-sharing website as opposed to sampling estate agents. We chose SpareRoom since it is the largest flat-sharing website in the UK, with millions of registered users.[8] It is also largely free to use, except for a minority of landlords

[8] https://www.spareroom.co.uk/content/about/about-spareroom-new/

who require 'early bird' access to contact.[9] It was therefore a natural choice for the experiment.

Ethical issues. One concern with such experiments is that it is impossible to ask subjects for informed consent in advance of their participation (Riach & Rich, 2002). Another concern is that sending fictitious applications wastes landlords' time and even creates the possibility that they might decline other (genuine) applicants whom they would otherwise accept. To minimise this risk, we promptly and politely declined all proposed viewings from interested landlords. While this did not eliminate the social costs of the experiment, we felt that the remaining costs were justified by the importance of the subject matter.

Choice of names. We selected names in accordance with two criteria. First, we wanted the names to be common to ensure a wide applicability of our results. Second, we wanted names to clearly signal ethnicity in the intended way, a key assumption on which the entire experiment rests.

In order to satisfy the first criterion, we chose English and Muslim names from the set of most common baby names in 2016.[10] (In hindsight, it would have been better to choose a previous year since the distribution of names today is determined by naming trends in previous decades.) For the English names, we chose Oliver, Charlie and Olivia, in addition to some other names which did not end up in the sample. These were all within the ten most popular baby names for their gender. For the Muslim names, we chose Mohammed, Ahmed, Fatima and

[9] Adverts are free to contact if they are more than 7 days old or the landlord has paid for the privilege. Conceivably contacting only free adverts could reduce the external validity of the results. However, the results are still valid for tenants who apply to free adverts, which in any case comprise approximately 4/5 of the adverts on SpareRoom.

[10] https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/babynamesenglandandwales/2016

Jamila (again in addition to other names which did not end up in the sample). These names were all within the ten most popular Muslim names for their gender.

To verify that these names would signal ethnicity in the intended way, we conducted a pre-experimental survey on Mechanical Turk. We restricted respondents to those within the UK, since it is UK perceptions which we were trying to measure. The results clearly show that respondents could 'correctly' categorise names as Muslim or non-Muslim. Not one respondent categorised any of the English names as Muslim. Meanwhile, every Muslim name was categorised as such by at least 6/7 of the sample. Table A1 in the Appendix provides the full results.

We also wanted to include some 'European' (but not English) names. This was to test whether any discrimination found against Muslims was 'merely' xenophobia, so would disappear when Muslim applicants were paired with European applicants. In the event, the only European names which ended up in the sample were Valérie and Francisca. To verify that these would not be seen as English, we carried out another pre-test on Mechanical Turk. As Table A1 reveals, the results were as expected. Valérie and Francisca were classified as non-English by over 90% of respondents, whereas no English name was classified incorrectly by more than 10% of respondents. As before, differences in classification rates were statistically significant between every English and non-English name (see A1 for the full results).

Finally, we decided to include some 'lower class' names. This was to investigate whether any discrimination found against Muslims was purely class-based, so would disappear when Muslim applicants were paired with lower class (English) applicants. To identify some names widely viewed as 'lower class' we began with two sources: the names used in a correspondence study of the effect of class in the

UK (Jackson, 2009) and a list of the least-represented names within the University of Oxford undergraduate body relative to their incidence in the population (Clark, Cummins, Hao & Vidal, 2015, p. 35). From these two sources, we selected the names Kevin, Gary, Connor, Jade, Stacey and Donna and asked UK respondents on Mechanical Turk to say which names they considered lower class. The 'winners' of this contest were Jade and Gary who were classified as lower class by 93% and 63% of respondents respectively (see Table A2).

In the end, the only lower-class name in the sample was Gary, who was paired with Mohammed. The relevant question was therefore whether Gary or Mohammed was seen as a more lower-class name. Accordingly, we conducted a final survey on Mechanical Turk which asked respondents this question. As shown by Table A3, 4/5 of the respondents chose Gary as the more lower-class name, validating our choice. Although the sample size was small, we could easily reject the hypothesis that each respondent was equally likely to choose Gary or Mohammed as the more 'lower class' name ($p < 0.01$).

Creation of accounts. In order to avoid sending a suspicious number of messages from any one account, we made each name four separate accounts on SpareRoom. To further reduce suspicion, all accounts were email verified and associated with a unique IP address using a private proxy service. All emails followed the format: firstname_randomnumber@gmail.com

Creation of messages. All messages had the same structure. The applicant explained that they had recently moved to the area and asked to see the property. To increase the salience of names, these were included in the body of the message as well as featuring in the email address displayed below the message box. We kept the messages as concise as possible to reduce random variation between

applicants. For instance, Mohammed and Gary sent the following pair of messages:

> I have just moved to London and was wondering if I could have a viewing?
> Mohammed

> My name is Gary and I recently moved to the area. Could I see the property?

Initially, the plan was to reuse each message twice, sending it once from a Muslim and once from a non-Muslim account. However, this turned out to trigger SpareRoom's spam filter. We therefore introduced minor variation between the two messages, such as replacing 'can' with 'could', and then assigned one message to a Muslim and one to a non-Muslim account using a random number generator. Hence, whenever a Muslim account (for instance) sent a message, a non-Muslim account sent almost (but not quite) the same message. For instance, Francisca's message was similar to Mohammed's and Fatima's message was similar to Gary's:

> I have recently moved to Leeds and was wondering if I can have a viewing?
> Francisca

> My name is Fatima and I just moved to the area. Can I see the property?

While these small changes do introduce random variation to the results, they seem unlikely to have made a significant difference. Furthermore, since the messages were randomly assigned, any variation that does exist does not introduce bias.

We felt that sending a given landlord multiple pairs of messages would look suspicious. We therefore decided to only message every landlord twice. In practice, we sometimes accidentally contacted landlords with multiple pairs, but all such landlords were dropped at the analysis stage.

Scope. We sent messages in the UK's four largest metropolitan areas: London, Manchester, Birmingham and Leeds-Bradford. In each city, we planned to apply to 560 properties, sending Muslim-English pairs and Muslim-European pairs in a 2:1 ratio. We planned to send all of the possible gender-matched pairs of names in equal numbers. When sending each pair, we planned to send one set of applications immediately after the other and to send the Muslim application first in exactly one-half of cases.

Unfortunately, the experiment did not go as planned for three reasons. First, SpareRoom's spam filter prevented many of the accounts from sending messages at all. Second, in many cases our accounts were simply suspended by SpareRoom. Third, even in some cases where the messages did get through, SpareRoom warned users that our messages were a hoax (something we learned from landlord responses). As time went on, SpareRoom became increasingly sophisticated at intercepting our messages and so we were forced to abort the experiment, having successfully sent 1,368 messages. While this reduced our sample, it did not introduce selection bias (unless SpareRoom's decision to suspend accounts somehow depended on how discriminatory a set of replies they generated.)

As it turned out, in the large majority of the pairs which sent successfully, the message from the Muslim had been sent first. Although one set of messages was always sent immediately after the other, and landlords were free to offer viewings to both applicants (as many did), this still may have helped the Muslim applicants.

Therefore, if one considers order effects to be important, one may wish to view our estimates as putting a lower bound on discrimination against Muslims.

Data collection and coding. We recorded landlord responses to each pair three days after that pair's messages had been sent. In addition to politely declining any offers of viewings, we also recorded the landlord's name, their 'type' (estate agent, current flatmate etc.), the property type (entire flat, house share etc) and the property price.

Messages were manually coded in one of three ways: an offer of a viewing ('yes'), questions without an explicit offer ('questions'), or no offer and no questions ('no'). We then used landlord names to identify their gender and whether they were Muslim. We identified name genders using an online database.[11] We determined whether a name was Muslim by checking whether the majority of people listed on that name's Wikipedia page came from Muslim majority countries. While this is not a perfect criterion, in practice it was always very clear whether to classify a name as Muslim (or not); so any reasonable criterion would have led to similar results.

Pre-registration. In the interests of transparency, the experiment and plan of analysis were pre-registered on the Open Science Framework.[12] With the one exception of the sample size, which turned out far lower than expected due to unpredicted actions by SpareRoom, the experiment and analysis conformed closely to the original plan.

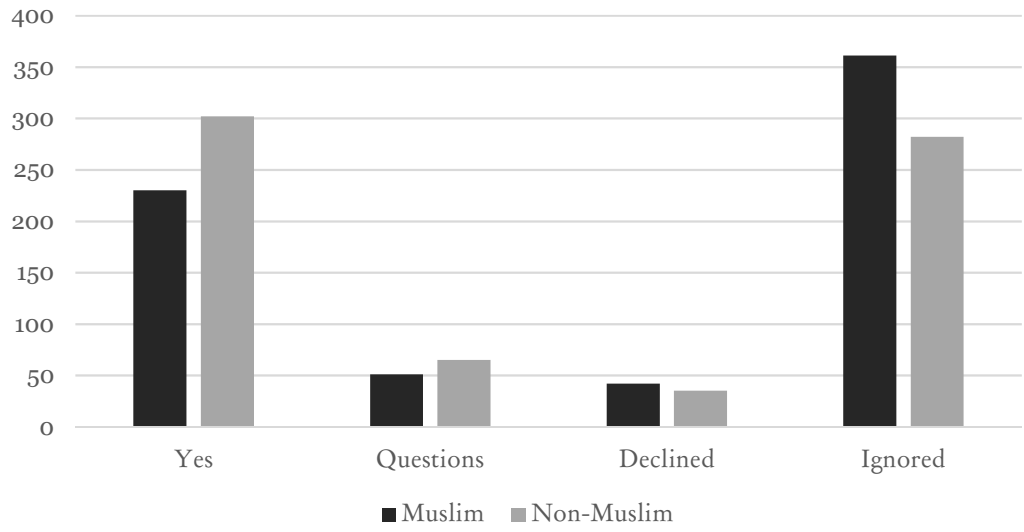
[11] http://genderchecker.com/

[12] https://osf.io/qc3zt/

5 Is There Discrimination?

To examine whether there is discrimination, we begin with the distribution of landlord replies (Table 1). As can be seen, most landlords treat both applicants in the same way, either accepting or (more likely) rejecting both. However, there is a sharp asymmetry amongst the remaining landlords in the sample. While 18 landlords accept the Muslim but reject the non-Muslim, 90 reject the Muslim but accept the non-Muslim. In other words, the non-Muslim is favoured in 83% of the discordant pairs. This disparity generates significant differences in the marginal distributions of responses, as shown by Figure 2. Muslims are less likely to get a 'yes', more likely to be declined, and far more likely to be ignored. Overall, Muslims are accepted in 33% of cases whereas non-Muslims are accepted in 44% of cases. This implies that non-Muslims are 32% (or 11 percentage points) more likely to receive a viewing.

Figure 2. Marginal Distributions of Responses



In Table 1, we categorise as response as 'yes' only if landlords explicitly suggest a viewing. However, one might wonder whether one obtains similar numbers

using an alternative definition of a 'yes'. Accordingly, we also consider a broad definition of 'yes', which includes landlords who ask questions but do not explicitly suggest a viewing. Table 2 reveals that unequal treatment remains substantial. Now, Muslims are accepted 41% of the time, whereas non-Muslims are accepted 54% of the time, implying that non-Muslims are 30% (or 13 percentage points) more likely to receive a viewing. (In future, we will use x/y to denote the number on the narrow definition and broad definition respectively.)

TABLE 1 - LANDLORD REPLIES

	Muslim	
Non-Muslim	Yes	No
Yes	212	90
No	18	364

Notes: Replies are counted as a 'yes' if landlords suggested a viewing.

TABLE 2 - LANDLORD REPLIES

	Muslim	
Non-Muslim	Yes	No
Yes	261	106
No	20	297

Notes: Replies are counted as a 'yes' if landlords suggested a viewing or asked the applicant questions.

Unequal treatment does not imply that any landlords have discriminated. After all, there are many non-discriminatory reasons why a landlord might reply to one applicant but not the other (for instance, some landlords may have only read one of the messages). Accordingly, we need to check whether the data can be accounted for by a model in which landlords randomly favour one applicant over another with equal probability (or otherwise reject or accept both). Carrying out this procedure is equivalent to conducting a binomial sign test. The logic of the test is straightforward: if the n landlords who chose one applicant over another did so randomly with equal probability, then the number who favour the Muslim is distributed binomially with n 'trials' and a 'success probability' of $1/2$.[13] This observation allows for straightforward calculation of p values (which we double

[13] By the symmetry of the binomial distribution, it is equivalent to consider the number who accept the non-Muslim but reject the Muslim.

to perform two-sided tests).[14] Using this (exact and non-parametric) test, we obtain the $p < 0.0001$ on either definition of 'yes' and therefore reject the hypothesis of no discrimination.

One might also wonder whether the results are driven by a particular applicant name. If a certain name were to have negative connotations which had nothing to do with religion, then that could generate misleading evidence of religious discrimination. Accordingly, we repeat the analysis, each time dropping one of the names from the sample. (Given the paired design, dropping one name always requires dropping at least one other.) As Table 3 reveals, the differences in acceptance rates remain fairly constant and p values remain essentially 0.

TABLE 3 - OMITTING NAMES

Omitted Names	Difference (narrow)	Difference (broad)
Ahmed, Oliver	0.11***	0.14***
Jamila, Olivia	0.11***	0.13***
Mohammed, Charlie, Gary	0.11***	0.11***
Mohammed, Charlie	0.11***	0.11***
Mohammed, Gary	0.10***	0.13***
Valérie, Fatima	0.11***	0.13***
Francisca, Fatima	0.10***	0.12***
Valerie, Francisca, Fatima	0.10***	0.12***

Notes: The narrow (broad) difference is the difference in acceptance rates between non-Muslim and Muslim applicants on a narrow (broad) definition of 'yes'. p values are calculated using a binomial sign test (otherwise known as an exact McNemar's test). ***denotes p < 0.001.

[14] To invert the test and so obtain confidence intervals, it is easiest to rely on the central limit theorem to deliver an asymptotic approximation to the distribution of a suitable transformation of the difference in acceptance rates (this is the McNemar test). In line with Agresti & Caffo (2000) we add four pseudo-observations to bring coverage probabilities closer to the target of 95%. (A 'coverage probability' is the probability that a confidence will contain the true parameter. This need not equal the nominal probability of a confidence interval, as consideration of a Bernouilli random variable with an estimated mean of zero indicates.)

In theory, discrimination could be driven by either differences in reply rates or differences in acceptance rates given a reply ('conditional acceptance rates'). To quantify this, we note that:

$$\mathrm{P}(\mathrm{Yes}) = \mathrm{P}(\text{ Yes} \mid \text{Response})\mathrm{P}(\mathrm{Response})$$

By the sample analogue of this fact, the ratio in acceptance rates is the ratio of response rates multiplied by the ratio of conditional acceptance rates. The former is 1.25, whereas the latter is just 1.05, indicating that differential response rates explain 5/6 of the discrimination. However, even considering just those landlords who reply to both, we can still find discrimination, albeit with higher p values than before ($p = 0.03/0.04$). Thus, although discrimination is driven mainly by differential reply rates, differential conditional acceptance rates also play a role.

6 Is There a Politeness Deficit?

While looking through the messages, it is easy to form the impression that landlords are more polite to non-Muslims than Muslims, even if they ultimately treat them 'the same' by accepting or rejecting both. For example, one landlord told the non-Muslim applicant that he would be 'delighted to show [her] around at [her] earliest convenience' but replied to the Muslim applicant with just one word: 'Sure'. While both of these replies are coded as a 'yes' in our dataset, there is nonetheless a clear 'politeness deficit' in the treatment of the Muslim applicant.

Anecdotes are no substitute for data, however, so we set about testing for a politeness deficit more formally. We measure politeness by checking whether

landlords thanked applicants, greeted them or used the word 'please'.[15] These words have been shown to especially well predict subjective assessments of politeness by recent work in computational linguistics and machine learning (Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec & Potts, 2013; Aubakirova & Bansal, 2016).[16] We then count the frequencies of these words within the subset of landlords who either accept or reject both applicants, thereby looking for a 'politeness deficit' that exists above and beyond the discrimination documented previously.[17]

As Table 4 reveals, we find some evidence of a politeness deficit, with non-Muslims around 15% (or 9 percentage points) more likely to be greeted ($p < 0.01$). This striking for two reasons. First, by considering just the landlords who did not discriminate (in the sense of rejecting one applicant but accepting the other), we are presumably looking at a relatively unprejudiced subset of the sample. Yet even this subsample is significantly less likely to greet the Muslim applicant. Second, as discussed previously, the message from the non-Muslim was usually sent after the message sent by the Muslim. As a result, one would think that most landlords, having only recently written a message to another applicant, would struggle to summon up similar enthusiasm for their new enquiry.

---

[15] Applicants were counted as *thanked* if the landlord used the words 'thanks' or 'thank you'. They were counted as *greeted* if the landlord began with the words 'hello', 'hi', 'dear', 'good morning', 'good afternoon' or 'good evening'.

[16] We also checked message length, finding that Muslims receive significantly shorter messages than non-Muslims. However, since we know of no prior work demonstrating a link between message length and politeness, we choose not to report these results.

[17] It would not be meaningful to compare the frequencies of these words for all landlords in the sample since the distribution of outcomes of replies is very different Muslim and non-Muslim applicants. For example, non-Muslim applicants are far more likely to get a reply (as previously noted) and this alone would tend to make landlords more likely to use any particular word when replying to non-Muslim applicants.

Table 4 - Politeness		
	Ratio (narrow)	Ratio (broad)
Greetings	1.15**	1.15**
Gratitude	1.04	1.03
Please	1.02	1.00

Notes: Ratio (narrow) and ratio (broad) provide the non-Muslim frequency divided by the Muslim frequency on the narrow and broad definitions of 'yes'. p values are calculated with a binomial sign test. ** denotes $p < 0.01$.

However, we also obtain some null results. While Muslims are slightly less likely to be thanked, this difference is not significant. Meanwhile, there are essentially no differences in the frequencies of the word 'please' (Table 4). Therefore, these results need to be interpreted cautiously and are not nearly as strong as our previous finding of discrimination.

Given that we find a politeness deficit on some dimensions but not all, it is natural to aggregate these dimensions into an overall politeness index. To do so, we first standardise each component by subtracting off its mean and dividing by its standard deviation. We then weight *greetings*, *gratitude* and *thanks* in a 0.43:0.49:0.87 ratio, the weights found by Danescu-Niculescu-Mizil et al. (2013) to optimally predict subjective assessments of politeness. (Reassuringly, similar weights are derived by Groenland, Oosterhuis, van Doorn and van Rozendaal in one of the few other papers on this topic.) Having constructed the index, we can compare how the Muslim and non-Muslim applicants score. We find not only that the Muslims score lower, as expected, but that the difference is statistically significant ($p = 0.02$).

Although the chosen weights do enjoy some justification, they may still appear rather arbitrary. Accordingly, we ask to what extent the results are robust to

different choices of weights. To make this question more concrete, we investigate how the weights can vary while retaining statistically significant differences on the index at the 95% level. If we use paired t test to calculate p values, then we are testing the hypothesis that the average difference in politeness equals zero. The (asymptotically normal) test statistic is simply:

$$t \;=\; \frac{\sqrt{n}\;\hat{d}}{\sqrt{\widehat{\mathrm{Var}[\hat{d}]}}} \;\rightarrow\; N[0,1]$$

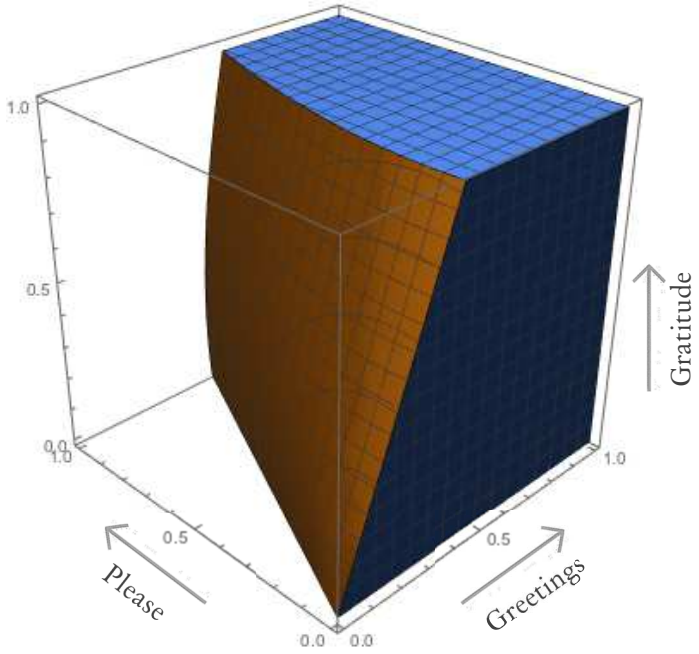
where $\hat{d}$ is estimated average difference in politeness as measured by the index, $\widehat{\mathrm{Var}[\hat{d}]}$ is its estimated variance and n is the sample size. The average difference in politeness is simply the weighted average of the average differences on each dimension of the index. That is, $\hat{d} = \sum_{i=1}^{3} \alpha_i \hat{d}_i$ where $\hat{d}_i$ is the estimated mean difference on dimension i (say *gratitude*) and α_i is its weight. Hence, we can write the test statistic as:

$$t \;=\; \frac{\sqrt{n}\;\sum_{i=1}^{3} \alpha_i \hat{d}_i}{\sqrt{\sum_{i=1}^{3} \alpha_i^2\, \widehat{\mathrm{Var}[\hat{d}_i]} + 2\sum_{1\leq i\leq j\leq 3} \alpha_i \alpha_j \widehat{\mathrm{Cov}(\hat{d}_i, \hat{d}_j)}}} \tag{7}$$

If we define α_1, α_2 and α_3 as the weights on the *greetings*, *gratitude* and *please* dimensions respectively, and calculate the estimated means, variances and covariances from the data, we see that the set of weights that give us a significant result are the values of α_1, α_2 and α_3 that satisfy:

$$\frac{\sqrt{237}\;[0.2\alpha_1 + 0.01\alpha_2 + 0.03\alpha_3]}{\sqrt{0.7\alpha_1^2 + 0.5\alpha_2^2 + 0.5\alpha_3^2 + 2(-0.03\alpha_1\alpha_2 + 0.2\alpha_1\alpha_3 + 0.02\alpha_2\alpha_3)}} \;\leq\; 1.96$$

The reader is free to substitute in her preferred weights to see if they satisfy this inequality. Perhaps more intuitively, the weights that deliver a significant result can be visualised as the shape below:



Unsurprisingly, as one increases the weight on *greetings* – the measure on which the politeness deficit is greater – one is more likely to find a statistically significant politeness deficit. Conversely, as one increases the weight on *please* – the measure on which there is no deficit – one moves away from the set of weights that deliver us a significant difference.[18]

[18] From (7), it is easy to see that the test statistic is invariant to multiplying each weight by a constant (so only relative weights matter, as one would expect). As a result, the plot has the 'fractal' property of looking exactly the same if we 'zoom out'. In other words, although we have plotted $0 < \alpha_i < 1$, the same picture would be produced by plotting $0 < \alpha_i < c$ for any $c > 0$.

As the plot reveals, a worrying fraction of the space is empty, suggesting that our finding is not robust to alternative choices of weights. To check this more formally, we calculate how much one can change one weight while holding the others fixed and retaining a significant result. This is a simple matter of allowing (7) to hold with equality, setting two of the weights equal to the preferred values and finding the third with the quadratic formula. The (positive) solutions are given in Table 5 and show the fragility of our result. For example, we need only change the weight on *greetings* from 0.43 to 0.38 to lose overall significance. Although these values assume the 'narrow' definition of 'yes', using the 'broad' yes produces almost identical results.

Table 5 - Is the Politeness Deficit Robust?

Dimension	Suggested weights	Critical Values
Greetings	0.43	> 0.38
Please	0.49	< 0.69
Gratitude	0.87	< 1.03

Notes: The critical values are calculated by setting two weights equal to their suggested values and finding the smallest deviation of the third weight from its suggested value that robs the politeness deficit of statistical significance.

Of course, deviations from our preferred set of weights need not be unilateral: all three weights could change at the same time. We therefore look for the shortest path from our set of preferred weightings to the boundary of the permissible set, using Euclidean distance as our metric of length. The problem is then to find the $\alpha_1, \alpha_2, \alpha_3$ which:

$$\text{minimise } \sqrt{(\alpha_1 - 0.43)^2 + (\alpha_2 - 0.49)^2 + (\alpha_3 - 0.87)^2}$$

subject to (7) holding with equality

The solutions are: $\alpha_1 = 0.38, \alpha_2 = 0.50, \alpha_3 = 0.88$.[19] Again, the deviations appear rather small and we only need decrease the weight on *greetings* by 12% to lose significance. This too highlights the fragility of our results. Thus, although we do find evidence of a politeness deficit, this evidence is at most suggestive and in need of replication from further research.

7 What Explains the Discrimination?

We now turn to the question of what explains the discrimination that we see in the data. We are unusually well placed to answer this question thanks our 'lower class' and 'European' controls, plus our data on the landlords whom we contacted and their properties. We begin with the question of class.

7.1 Is the Discrimination Class-Based?

In an influential paper, Fryer and Levitt (2004) show that distinctively African-American names are a strong signal of low socio-economic status, which suggests that what looks like racial discrimination might actually be class-based discrimination. Similarly, one might suggest that the discrimination we uncover against Muslims is driven not by an aversion to their ethnicity or religion but rather by the belief that they are poorer and/or lower class. This possibility needs to be taken seriously in light of research demonstrating the existence of discrimination against lower class applicants in the UK labour market (Jackson, 2009).

If the discrimination against applicants with Muslim names were purely class-based, then it would disappear in the pair in which a Muslim name was matched

[19] To solve this for some variable (say, α_1) we can use (7) and the quadratic formula to express α_1 in terms of α_2 and α_3, substitute into the objective function and set all 3 partial derivatives equal to 0. This yields three simultaneous equations whose solution is the minimum distance deviation for α_1.We then repeat for α_2 and find α_3 with the constaint.

with a lower-class English name (Gary/Mohammed). However, this is not what we see in the data. Gary is 11-12 percentage points more likely to be offered a viewing than Mohammed; a gap which closely resembles the overall difference in acceptance rates in the sample. Although Gary and Mohammed sent just 186 applications between them, the difference in their acceptance rates is statistically significant ($p < 0.01$). Hence, we can reject the suggestion that the discrimination is purely class-based; and our point estimates provide no evidence that it is even partially class-based.

7.2 Is the Discrimination 'Just' Xenophobia?

Next, we investigate whether discrimination against Muslim applicants occurs simply because they are not English. To test this, we examine whether discrimination also occurs against Muslim applicants when they are paired with European applicants. This requires restricting intention to the Valérie/Fatima and Francisca/Fatima pairs. We find that discrimination persists within this subsample, with the European applicants 14 percentage points more likely to be offered a viewing than the Muslim applicant. This is actually a slightly larger difference than usual, although the difference is not statistically significant. Moreover, the difference between European and Muslim applicants is statistically significant ($p < 0.01$) even though these pairs comprise just 16% of the sample. It follows that discrimination against Muslims is more than 'just' xenophobia.

7.3 Is the Discrimination 'Statistical'?

Even though the discrimination seems to be driven by the fact that applicants have Muslim names, this does not imply that it is driven by a dislike of Muslims. As explained in Section 2.2, an alternative explanation is that landlords rationally use Muslim names as a signal that applicants are less likely to pay their bills on time, keep the property in good condition and so on – in other words, that these

tenants are of lower 'quality'. Before testing this explanation with our data, we first provide some independent reasons to be sceptical.

First, despite countless papers which discuss models of statistical discrimination, there is a complete absence of empirical evidence that decision-makers form rational beliefs about demographic groups. In fact, this goes against the consensus in social psychology, which maintains that stereotypes typically inflate existing differences between groups and may reflect outdated information (Pager & Karafin, 2009), even if they typically exhibit some correlation with underlying reality (Jussim, Crawford & Rubinstein, 2015).[20]

Models of statistical discrimination also ignore the large body of literature showing that people are poor intuitive statisticians. To provide just a few examples, people make estimates that are swayed by informationally irrelevant anchors, form confidence intervals that are far too narrow, and find conditional probabilities especially hard to calculate, frequently confusing the probability of A given B with the probability of B given A (Alpert and Raiffa, 1969; Kahneman and Tversky, 1974). This provides yet another reason to be sceptical of statistical discrimination models.

Finally, we are particularly sceptical of the idea that people form rational beliefs when it comes to politically sensitive topics such as Islam. To verify this intuition, we asked UK respondents on Mechanical Turk to estimate the percentage of successful, failed or foiled terrorist attacks in the UK between 2005 and

[20] Although certain stereotypes (including national stereotypes and political stereotypes) are widely considered to lack any predictive power whatsoever, there is some debate about the 'accuracy' of other stereotypes. While Jussim et al. show that these typically exhibit low but positive correlations with underlying group differences, this only demonstrates that they contain a grain of truth, not that they are rationally warranted. See Bian and Cimpian (2017) for a related objection.

2016 inclusive that were conducted by individuals who professed Islamic goals. According to data from the Global Terrorism Database, the correct answer is around 2%.[21] However, as Table A4 reveals the mean answer was around 43%, with only two of the 49 respondents guessing the correct answer or lower. This is striking since starting the period in 2005 (a year with an unusually large number of Muslim terrorist attacks) made it as hard as possible for respondents to overestimate the true figure. Nonetheless, popular beliefs bore very little relation to reality, a result that sits uncomfortably aside models of rational statistical discrimination.

Having provided some general arguments against statistical discrimination, we now turn to the data from our experiment. The logic of our first test is straightforward. In standard models of statistical discrimination, all agents have the same beliefs. Indeed, formally speaking, agent homogeneity is what distinguishes them from taste-based models (Aigner & Cain, 1977). We thus examine whether there is any difference between Muslim and non-Muslim landlords when it comes to discrimination.

The results are striking. The 55 landlords with at least one Muslim name actually accept slightly more Muslim than non-Muslim applicants (23/25 *vs.* 22/25). Therefore, in contrast to the rest of the sample, there is no evidence whatsoever that this group are discriminating against Muslims.

Although there appears to be a difference between Muslim and non-Muslim landlords, this might not be statistically significant. Accordingly, we regress the outcome of the application on the ethnicity of the applicant's name, the ethnicity

[21]https://www.start.umd.edu/gtd/search/Results.aspx?start_yearonly=2005&end_yearonly=2016&start_year=&start_month=&start_day=&end_year=&end_month=&end_day=&country=603&asmSelect1=&dtp2=all&success=yes&casualties_type=b&casualties_max=

of the landlord, the product of these variables.[22] By a well-known argument, the hypothesis that Muslim applicants discriminate less is equivalent to the hypothesis that the coefficient on the interaction term is negative. Since each landlord features twice in the sample, violating independence across observations, we cluster standard errors at the landlord level (in addition to using heteroskedasticity robust standard errors).

The first column of Table 6 summarises the results. The coefficient on 'Applicant is Muslim' underlines what we know already, namely that landlords discriminate against Muslim applicants. More interesting, however, is the coefficient on the interaction term. This is slightly larger than the coefficient on the Muslim dummy, suggesting that Muslim landlords are (if anything) more likely to accept Muslim than non-Muslim applicants. The coefficient of the interaction term is highly significant ($p < 0.01$), suggesting that Muslim landlords really do discriminate less. As Column (3) shows, this remains true on the broader definition of a 'yes'.

Since the religion of landlords is not randomly assigned, Muslim landlords may differ from non-Muslim landlords in systematic ways that in turn determine the probability of discrimination. For example, it might be that the Muslim landlords offer more expensive properties and landlords who offer more expensive properties are less likely to discriminate. We attempt to control for such effects by including not merely covariates but also their *interactions* with the Muslim

[22] We report results from a linear probability model as opposed to probit or logit models since non-linear models present certain difficulties in the presence of interaction terms. In order to make probit or logit models interpretable, researchers typically report marginal effects but these are not equal to interaction effects in probit or logit models, and need not even have the same sign (Ai & Norton, 2003). In any case, in unreported results, we find that logit and probit models produce extremely similar results (in terms of statistical significance) to the linear probability model.

applicant dummy (since these interactions capture the effect of these characteristics on discrimination). As Table 6 shows, adding these variables makes essentially no difference to the interaction coefficient, a result that holds on either definition of 'yes'. Although this suggests that our result is not driven by omitted variable bias, our lack of detailed host and property data makes this hard to say for certain.

TABLE 6 - DO MUSLIM LANDLORDS DISCRIMINATE LESS?

	(1)	(2)	(3)	(4)
Constant	0.45***	0.68**	0.54***	0.72***
Applicant is Muslim	-0.12***	-0.05	-0.14***	-0.1***
Landlord is Muslim	-0.05	-0.09	-0.08	-0.13*
Interaction	0.13***	0.15***	0.14**	0.15**
Covariates?	No	No	Yes	Yes

Notes: Columns (1) and (2) report coefficients from a regression of a 'yes' response on applicant religion, landlord religion and their interaction. Columns (3) and (4) include the interactions of host and property covariates with the religion of the applicant. Columns (1) and (3) use the 'narrow' definition of 'yes'; columns (2) and (4) use the 'broad' definition. * denotes $p < 0.05$, ** denotes $p < 0.01$ and *** denotes $p < 0.001$.

One might wonder whether these results are driven by our rather permissive classification of landlords as Muslims. As stated earlier, we classified a landlord as Muslim if at least one of their names was Muslim/Arabic. However, this includes landlords with one Muslim and one non-Muslim name (a fictitious example would be John Ali). We therefore check whether the results are robust to classifying such landlords as non-Muslim. We find slightly stronger results than before, with p values below 1% in all specifications.

Although this result is inconsistent with the standard model of statistical discrimination, it is conceivably consistent with a model with heterogenous beliefs. Specifically, one might claim that Muslim and non-Muslim landlords are, while rational and non-prejudiced, nonetheless unaware of the true (religion-specific)

distributions of tenant quality. As a result, they need to form estimates using the limited information available to them. The claim would then need to be that the kind of information to which Muslim landlords have access leads them to treat Muslim applicants more favourably than non-Muslim landlords do.

While this is a logical possibility, there is actually no reason to suppose that the information to which Muslim and non-Muslim applicants have access varies in the required way. To be sure, Muslim landlords likely have *more* information about Muslim applicants since they may (on average) know more Muslims. However, more information does not mean more positive information. Assuming that all landlords are estimating religion-specific means using sample means, an unbiased estimator, then the fact that they have a greater sample size does not suggest that their assessments of Muslim quality will be more favourable.

Similar remarks apply to variances. Even though Muslim landlords likely have access to a larger sample of Muslims, this does not mean that their estimates will tend to be lower, at least if they are using an unbiased estimator. We belabour this point because it does not seem to be appreciated by the literature on statistical discrimination. For example, Bertrand and Duflo (2017) claim that:

> Limited de-facto contact between in-group and out-group members will imply that majority employees or co-workers will be fairly ignorant about the quality of minorities . . . which, in the presence of risk aversion, will also trigger more statistical discrimination. (p. 5)

Actually, all we can say is that limited contact increases the variance of the estimators (i.e. standard errors), not that it increases point estimates of variances. It is presumably this subtle distinction – between the estimated variance and the variance of an estimator – which underlies this common misconception.

To clarify, although Muslim landlords do not have access to more positive information about Muslims, they may still have more positive beliefs. Indeed, decades of research in social psychology demonstrates that people typically have positive views about members of their own social and ethnic groups (Mullen, Brown & Smith, 1992), a general rule to which Muslims are unlikely to prove an exception. This notwithstanding, we see no reason to suppose that any differences in beliefs between Muslim and non-Muslim landlords are rationally warranted by the evidence. In the language of Section 2.2, we are allowing for the possibility of 'error discrimination' while arguing against the possibility of statistical discrimination.

There is a caveat, however. By the very fact that people tend to favour their ingroup, it might be that Muslim tenants behave in a 'better' way towards Muslim landlords than their non-Muslim counterparts. In that case, Muslim landlords really are dealing with different distributions of applicant quality than non-Muslim landlords and their differential discrimination rates might simply be rational responses to this disparity in distributions. This is not our preferred explanation: we find it hard to believe that many tenants decide how frequently to pay their rent (for example) based on the ethnicity of their landlord. However, this is not an explanation that we can reject with our data.

7.4 Is the Discrimination 'Taste-Based'?

As discussed in Section 2.2, an alternative explanation for the discrimination is that landlords dislike interacting with Muslim applicants. Unlike statistical discrimination, this theory also gracefully explains the previous finding that Muslim landlords do not discriminate, assuming that Muslim landlords like interacting with Muslim applicants more than non-Muslim landlords do.

To test whether discrimination depends on interaction (Becker typically uses the word 'contact'), we begin by comparing properties which are offered as a whole with properties which are shared. If discrimination were interaction based, one would expect discrimination to be greater in the shared properties. As before, we need to account for the fact that property type is not randomly assigned, so might vary with other factors that influence discrimination rates. Accordingly, we model the probability that an applicant is offered a viewing as a function of applicant ethnicity, property type (shared or not), the interaction of the two and all covariates and their interactions with the applicant ethnicity dummy. We cluster standard errors at the landlord level and used heteroskedasticity robust standard errors. We carry out the exercise twice, once on the narrow definition and once on the broad definition of 'yes'.

Appendix Table A5 presents the results. Surprisingly, the coefficient on the interaction term is negative in all specifications, suggesting that there is less discrimination in shared properties. However, this is not remotely significant, with all p values above 49%. Furthermore, although the point estimates are negative, 95% confidence intervals always include positive values. Therefore, while our data offers no support for the Becker model, it does not offer strong evidence against it either (largely because there are just 43 non-shared properties in the entire sample).

We next investigate how discrimination varies with landlord type. Fellow house or flatmates have far more interaction with new tenants than live out landlords or estate agents, although both types of 'landlord' may still show the prospective tenant around. The Becker model therefore predicts greater discrimination from house or flatmates (who will be living with the applicant) than from the other types of respondent in our sample.

To test this, we use the same estimation procedures as before on both definitions of 'yes'. Table A6 provides the results. As before, the coefficients on all the interaction terms are negative, suggesting that live out landlords and estate agents discriminate more by about 4-5 percentage points. Since there we have many more such landlords than entire (non-shared) properties, p values are now far lower, ranging from 10% - 29%. However, these results remain insignificant by conventional criteria and confidence intervals for the interaction coefficient continue to contain positive values (the prediction of the Becker model). Hence, while our data once again offers no support whatsoever to the interaction hypothesis, the size of our sample makes it hard to reject it entirely.

Furthermore, it is possible to think of ways in which the Becker model might be consistent with our data. One idea is that landlords dislike interacting with Muslims for any amount of time at all; but if they are going to interact with Muslims they are totally indifferent to the amount of time for which they need to do so. Since all types of landlord in our sample need to interact with applicants to some extent (if only to show them the property), this would explain why we do not find a statistically significant difference in acceptance rates across landlord types. However, while logically possible, it is hard to think of how one might motivate such a utility function. One would have thought that anything that would make landlords dislike interacting with Muslims for a short amount of time (such as a perceived lack of common interests) would make longer interactions even worse.

More plausibly, perhaps, one might claim that live-out landlords act in accordance not with their own ethnic preferences but rather in accordance with those of their tenants (or what they take these preferences to be). If live out landlord and estate agents believe that their tenants dislike interaction with Muslims, then they may themselves discriminate against Muslims so they can charge higher

prices (live out landlords), retain tenants as customers (estate agents) or ensure that all vacant rooms in the house are filled (both). This could also explain why we do not see a statistically significant difference between types. While we cannot reject this explanation without data, the relatively sophisticated reasoning which it assumes would appear to be at odds with evidence that discrimination is driven by unconscious preferences and beliefs (Bertrand, Chugh & Mullainathan, 2005; Rooth, 2007).

Whether or not these results of this section constitute strong evidence against the Becker model, they certainly have legal implications. According to Schedule 5 of the Equality Act 2010, it is legal for a landlord to discriminate on the basis of religion (but not race) if they will be sharing 'small premises' with the tenant. Hence, unless the discrimination against Muslims were racial rather than religious in nature – which would seem difficult for a court to establish – the discrimination uncovered by current flatmates and housemates would be legal. However, this section demonstrates that estate agents and live-out landlords are (if anything) more likely to discriminate than those who would live with the applicant. Indeed, we can easily reject the hypothesis of no discrimination ($p < 0.01$) for just this subsample. Therefore, some of the estate agents and live-out landlords in our sample are behaving illegally, even if discrimination by current flatmates is legally protected.

7.5 Other Explanations

Our data do not support either of the leading theories of discrimination. On the one hand, discrimination entirely disappears once we restrict attention to Muslim landlords, a difference which is statistically significant ($p < 0.01$). This is inconsistent with standard models of statistical discrimination and hard to reconcile with even more permissive models which allow landlord beliefs to vary. On the other hand, the landlords who interact more with the tenant actually

discriminate less, in apparent contradiction of Becker's (1957) model of taste-based discrimination. We therefore ask what does explain the discrimination which we uncovered.

Although we cannot definitively answer this question, we can offer some suggestions. One idea is that discrimination is generated by irrational beliefs about the quality of Muslim and non-Muslim applicants (so-called error discrimination). This would explain why discrimination is not driven by interaction *per se*, but nonetheless disappears once we restrict attention to Muslim landlords. The only auxiliary assumption on which this explanation relies is that Muslim landlords have more positive beliefs about Muslim applicants than non-Muslim landlords do, an assumption that receives overwhelming support from research in social psychology (Mullen et al., 1992 present a meta-analysis).

Another idea is that discrimination occurs because landlords enjoy helping members of their own group (or enjoy harming members of other groups). This contention also enjoys considerable support from research in social psychology and other disciplines.[23] Since most landlords in the sample are non-Muslim, this would explain why we observe discrimination. However, it would also explain why Muslim landlords do not discriminate against Muslims applicants.[24] Furthermore, it is consistent with the apparent unimportance of interaction: agents

[23] Lane (2016) summarises the evidence from experimental economics, showing (for example) that participants in dictator games often donate more to members of their own ethnic group even though they have no financial incentive to do so. Balliet, Wu and Dreu (2014) summarise the evidence for 'ingroup favouritism' from the social psychology literature. The existence of ethnic conflict also provides evidence for this phenomenon.

[24] This explanation suggests that Muslim landlords should actually discriminate in favour of Muslim applicants. However, given the small number of Muslim landlords in our sample, we cannot reject this hypothesis with our data.

can enjoy helping members of their own group even if they do not interact with them.

We stress that while these explanations are not interaction based, they are close in spirit to Becker's notion of taste-based discrimination. Indeed, formally speaking they are completely consistent with the Becker model. All Becker needs is that employers (for example) act as if they are willing to pay a price to avoid hiring a minority group. Whether this stems from interaction, irrational beliefs or ethnic homophily is beside the point. Indeed, having reinterpreted the discrimination coefficient in the Becker model (as a measure of biased beliefs or homophily), one can proceed to derive all the classic results: that discrimination depends only on the biased beliefs/homophily of the 'marginal discriminator', that it increases in the relative size of the minority group, and so on (Becker, 1957).

8 Extensions

8.1 Exploratory Analysis

Although we were not particularly interested in these variables, we also estimated models with interaction terms to investigate how discrimination varies by city, landlord gender, applicant gender, landlord and property type (distinguishing on a finer basis than in Sections 7.3 and 7.4). As Table A7 reveals, none of these variables are significant at even the 10% level. Discrimination occurs for all genders, cities, landlord and property types, underscoring the robustness of our previous finding of discrimination.

8.2 Multiple Hypothesis Tests

One problem with testing multiple hypotheses (as in this study) is that it increases the probability of wrongly rejecting at least one hypothesis even if all the hypotheses are true. In other words, it increases the probability of a Type I

error. Since the null hypotheses in this study are zero effects, this inflates the probability that we will spuriously uncover a 'significant' result.

In order to control the probability of a Type I error, we can apply a Šidàk correction. (This is the exact version of the more popular Bonferroni correction.) Assuming independence of tests, the probability $\tilde{\alpha}$ of falsely rejecting at least one true null hypothesis given that all n null hypotheses are true is a simple function of each test's individual significance level α:

$$\tilde{\alpha} = 1 - (1 - \alpha)^n$$

As outlined in our pre-registration plan, the study was designed to test four main hypotheses:

1. Discrimination exists against Muslim applicants.
2. Even those landlords who do not formally discriminate are nonetheless less polite to Muslim applicants.
3. Muslim landlords are less likely to discriminate ('ethnic homophily').
4. Live-in landlords are more likely to discriminate ('interaction').[25]

To be sure, we conducted far more than four statistical tests in this paper. However, these tests were typically far from independent. For example, we conducted each test twice, once on either of our two definitions of 'acceptance'. These two sets of hypotheses were almost identical and always yielded extremely similar results, so it would not make sense to include them in the correction. In

[25] We also planned to test a hypothesis about a follow-up survey conducted on the landlords whom we contacted. However, it was not possible to conduct this survey due to time constraints.

fact, doing so would perversely discourage the kind of robustness checks carried out in the study.

We therefore believe that $n = 4$ is an appropriate correction (and certainly more conservative than the usual practice of making no correction whatsover). We thus calculate corrected p values using:

$$\tilde{p} = 1 - (1 - p)^4$$

As can be seen from Table 7, applying the correction does not broadly change the central messages of the study. There continues to be strong evidence that landlords discriminate against Muslims, but that Muslim landlords discriminate less ('ethnic homophily'). The interaction effect remains insignificant. However, the finding of a politeness deficit (which was already sensitive to the choice of weights) loses significance at the 5% level. This gives us a second reason to treat this finding as merely suggestive.

TABLE 7 - ŠIDÀK CORRECTION

Hypothesis	*p* values	Corrected values
Discrimination	0.00/0.00	0.00/0.00
Politeness	0.02/0.02	0.08/0.08
Homophily	0.00/0.00	0.00/0.00
Interaction	0.50/0.63	0.93/0.98

Notes: in each column, the first value uses the 'narrow' definition of 'yes' and the second uses the 'broad' definition. Although *p* values for 'homophily' and 'interaction' are reported for the simplest specification, including covariates changes little.

9 Conclusion

According to its statement on discrimination, SpareRoom recommends taking people "on the basis of who they are, rather than which 'boxes' they tick".[26] It turns out, however, that not all of its users follow this advice. Many landlords exhibit a strong preference for non-Muslim over Muslim applicants, an extremely robust result that persists across genders, cities, and landlord/property types. Overall, we estimate that non-Muslim applicants are between 30% and 32% more likely to be offered a viewing than otherwise identical Muslim applicants.

Neither of the two leading theories appear capable of explaining the discrimination. On the one hand, Muslim landlords do not discriminate, a result which is hard to reconcile with statistical models of discrimination. On the other hand, discrimination does not increase with interaction, in apparent contradiction with Becker's (1957) hypothesis. In contrast, either biased beliefs or ethnic homophily are able to simultaneously explain all of the findings in this study.

Although this study presents some strong and novel results, it also has many limitations. We begin with a discussion of some problems which afflict correspondence studies in general and then proceed with a discussion of some particular issues that this study faces.

First, we are unable to estimate the extent to which Muslim applicants are harmed by the discrimination we identify. If messages are almost costless to send and every type of property is offered by multiple landlords (i.e. within each category housing is a homogenous good) then the cost of discrimination will be

[26] https://www.spareroom.co.uk/content/default/discrimination/

small. Muslims are less likely to receive viewings but can easily offset this by increasing the number of applications they make. If, however, applications are either expensive to make or there are few houses within each category, discrimination will typically deny Muslim applicants their first choice of property. Hence, the welfare costs of discrimination depend on various factors above and beyond the difference in acceptance rates across groups.

Second, we are unable to identify the fraction of landlords who prefer non-Muslim applicants to otherwise identical Muslim applicants. In fact, this is a universal if unacknowledged problem with audit and correspondence studies. To see this, consider the joint distribution of responses, this time expressed in terms of percentages:

TABLE 8 - LANDLORD REPLIES

	Muslim	
Non-Muslim	Yes	No
Yes	31%	13%
No	3%	53%

Notes: Replies are counted as a 'yes' if landlords suggested a viewing.

The problem is that multiple models could account for the above data. One option is that 13% of landlords *never* accept Muslims, 3% *only* accept Muslims and the rest do not discriminate. In that case, 13% have a clear preference for non-Muslim applicants (and 3% have the opposite preference). However, landlords might be more likely to reject Muslim applicants even though they do sometimes accept them. For example, it might be that every landlord accepts both applicants with probability .31, rejects both with probability .53, accepts only the non-Muslim with probability .13 and accepts only the Muslim with probability .03. In that case, *all* landlords are biased against Muslims since they are more likely to choose the non-Muslim over the Muslim than do the reverse. Hence, we

cannot identify the fraction who prefer non-Muslim applicants. However, we can show that the fraction must be at least 10%.[27]

Third, it is not clear that the discrimination we uncover has any policy implications. One proposal to combat the discrimination might be to anonymise all applications. However, this proposal might be hard to implement (since it might undermine trust on the platform) and could actually harm Muslim users of SpareRoom. The current system allows landlords to discriminate against Muslims at the first opportunity. While this is regrettable, it might be better than a system which conceals applicant identities until the viewing stage, forcing Muslim applicants to incur even greater search costs by the time that they are ultimately rejected by prejudiced landlords.

There are also some more specific limitations with this study. First, as noted previously, due to exogenous interference, the Muslim applicants in the sample ended up substantially more likely to send the first message than the non-Muslim applicants. Although one set of messages was sent immediately after the first and landlords were free to accept both applicants (as many did), this may have given the Muslim applicants a small edge, meaning that our results actually understate the discrimination they face. Second, it is possible that online services such as SpareRoom are not representative of the rental market as a whole. However, given that discrimination is practised by both estate agents and landlords alike,

[27] A proof is as follows. In order to find the minimum possible percentage who prefer non-Muslims, we can restrict attention to models in which those who discriminate against Muslims do so with probability 1. Now, if this fraction is strictly less than 10%, the model needs to contain strictly more than 6% of landlords who randomly choose between the Muslim and non-Muslim with equal probability. Otherwise, it could not account for the additional 3% of landlords who reject the Muslim. However, the model would then predict that strictly more than 3% accept the Muslim but reject the non-Muslim, which is greater than the observed sample proportion.

we see little reason to countenance this suggestion. Third, while our sample is more than large enough to establish the existence of discrimination, we encounter some difficulties when analysing subsamples (notably when examining the effect of interaction). Again, this difficulty resulted from the premature ending of the experiment following action by SpareRoom, which meant that the sample was just over a quarter of the size originally planned.

The study also highlights opportunities for future research. The finding that discrimination does not increase with interaction is (to our knowledge) entirely novel but also rather weak on statistical grounds. It therefore stands in urgent need of replication. Future studies could also attempt to replicate the finding that Muslim landlords are less likely to discriminate against Muslim applicants, a hypothesis which has only been addressed by two previous studies which reach conflicting results (Ahmed & Hammarstedt, 2008; Carlsson & Eriksson, 2014).[28]

In addition, future studies could continue the study of 'politeness discrimination'. While at least one existing study examines differential politeness (Hebel et al., 2002), pairing a politeness analysis with an analysis of simple acceptance rates is far from standard in the literature. This is surprising since, once one has conducted a correspondence study, conducting a textual analysis (perhaps using the techniques proposed in this paper) is then straightforward. While this could be feature of future correspondence studies, examining differential politeness using data from the many correspondence studies that already exist would also be a worthy research topic.

[28] As noted earlier, these studies investigate whether landlords with 'ethnic minority' names are less likely to discriminate against Muslims. Hence, this is the first study examine whether specifically *Muslim* landlords are less likely to discriminate against Muslim applicants: the claim implied by the general hypothesis of 'ethnic homophily'.

10 Bibliography

Acquisti, A. & Fong, C. M. (2015). An Experiment in Hiring Discrimination Via Online Social Networks. SSRN: http://dx.doi.org/10.2139/ssrn.2031979.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Hoboken, New Jersey: Wiley.

Agresti, A., & Caffo, B. (2000). Simple and Effective Confidence Intervals for Proportions and Differences of Proportions Result from Adding Two Successes and Two Failures. *The American Statistician, 54*(4), 280-288.

Ahmed, A. M. & Hammarstedt, M. (2008). Discrimination in the Rental Housing Market: A Field Experiment on the Internet. *Journal of Urban Economics*, *64*(2), 362-372.

Ahmed, A. M., Andersson, L., & Hammarstedt, M. (2010). Can Discrimination in the Housing Market Be Reduced by Increasing the Information about the Applicants? *Land Economics, 86*(1), 79-90.

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80*(1), 123-129.

Aigner, D., & Cain, G. (1977). Statistical Theories of Discrimination in Labor Markets. *Industrial and Labor Relations Review, 30*(2), 175-187.

Alpert, M., and H. Raiffa (1969). A Progress Report on the Training of Probability Assessors. *Unpublished Report.*

Altonji, J. & Blank, R. (1999). Race and Gender in the Labor Market. In O. Ashenfelter & D. Card, (Eds.), *Handbook of Labor Economics, Vol. 3C* (pp. 3143–3259). Elsevier, North Holland.

Arrow, K. (1973). The Theory of Discrimination. In O. Ashenfelter & A. Rees (Eds.), *Discrimination in Labor Markets* (pp. 3-33). Princeton University Press.

Aubakirova, M., & Bansal, M. (2016). Interpreting neural networks to improve politeness comprehension. arXiv preprint arXiv:1610.02683.

Ayres, I., & Siegelman, P. (1995). Race and Gender Discrimination in Bargaining for a New Car. *The American Economic Review, 85*(3), 304-321.

Baert, S., Cockx, B., Gheyle, N., & Vandamme, C. (2013). *Do Employers Discriminate Less if Vacancies are Difficult to Fill? Evidence from a Field Experiment* (CESifo Working Paper Series No. 4093). Munich: CESifo Group Munich.

Baldini, M., & Federici, M. (2011). Ethnic Discrimination in the Italian Rental Housing Market. *Journal of Housing Economics, 20*(1), 1-14.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. Psychological Bulletin, 140(6), 1556.

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016). Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition. *American Economic Review*, *106*(6), 1437-1475.

Becker, G., (1957). *The Economics of Discrimination* (1st ed.). Chicago: University of Chicago Press.

Bengtsson, R., Iverman, E., & Hinnerich, B. T. (2012). Gender and Ethnic Discrimination in the Rental Housing Market. *Applied Economics Letters, 19*(1), 1-5.

Bertrand, M., Dolly Chugh, & Mullainathan, S. (2005). Implicit Discrimination. The American Economic Review, 95(2), 94-98.

Bertrand, M. & Duflo, E. (2017). Field Experiments on Discrimination. In E. Duflo and A. Banerjee, (Eds.), *Handbook of Field Experiments, Volume 1*. North Holland.

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review, 94*(4), 991-1013.

Bian, L., & Cimpian, A. (2017). Are stereotypes accurate? A perspective from the cognitive science of concepts. *Behavioral and Brain Sciences, 40*, 22-24.

Blommaert, L., Coenders, M., & Van Tubergen, F. (2014). Discrimination of Arabic-Named Applicants in the Netherlands: An Internet-Based Field Experiment Examining Different Phases in Online Recruitment Procedures. *Social Forces, 92*(3), 957-982.

Bosch, M., Carnero, M. A., & Farré, L. (2010). Information and Discrimination in the Rental Housing Market: Evidence from a Field Experiment. *Regional Science and Urban Economics, 40*(1), 11-19.

Carlsson, M., & Eriksson, S. (2014). Discrimination in the Rental Market for Apartments. *Journal of Housing Economics*, *23*, 41-54.

Carlsson, M., Fumarco, L., & Rooth, D. (2014). Does the Design of Correspondence Studies Influence the Measurement of Discrimination? *IZA Journal of Migration*, *3*(1), 11.

Carpusor, A. G., & Loges, W. E. (2006). Rental Discrimination and Ethnicity in Names. *Journal of Applied Social Psychology, 36*(4), 934-952.

Charles, K., & Guryan, J. (2008). Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination. *Journal of Political Economy, 116*(5), 773-809.

Clark, G., Cummins, N., Hao, Y. & Vidal, D. D. (2015). Surnames: A New Source for the History of Social Mobility. *Explorations in Economic History*, *55*, 3-24.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Edelman, B., Luca, M. & Svirsky, D. (2017). Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. *American Economic Journal: Applied Economics*, *9*(2), 1-22.

England, P. (1992). *Comparable Worth: Theories and Evidence*. New York: Aldine de Gruyter.

Fryer, R., & Levitt, S. (2004). The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics, 119*(3), 767-805.

Groenland, C., Oosterhuis, H., van Doorn, J., & van Rozendaal, T. Classification and visualisation of politeness. http://www.tivaro.nl/papers/NLP-1.pdf

Hebl, M. R., Foster, J. B., Mannix, L. M., & Dovidio, J. F. (2002). Formal and Interpersonal Discrimination: A Field Study of Bias Toward Homosexual Applicants. *Personality and Social Psychology Bulletin*, *28*(6), 815–825.

Heckman, J. (1998). Detecting Discrimination. *Journal of Economic Perspectives*, *12*(2), 101-116.

Jackson, M. (2009). Disadvantaged through discrimination? The role of employers in social stratification. *The British Journal of Sociology*, *60*(4), 669-692.

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (In)Accuracy in Perceptions of Groups and Individuals. *Current Directions in Psychological Science, 24*(6), 490-497.

Kaas, L., & Manger, C. (2011). Ethnic Discrimination in Germany's Labour Market: A Field Experiment. *German Economic Review*, *13*(1), 1-20.

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, *22*, 103–122.

Neumark, D. (2012). Detecting Discrimination in Audit and Correspondence Studies. *Journal of Human Resources*, *47*(4), 1128-1157.

Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2015). Racial Discrimination in the Labor Market for Recent College Graduates: Evidence from a Field Experiment. *The B.E. Journal of Economic Analysis & Policy, 15*(3).

Pager, D., & Karafin, D. (2009). Bayesian Bigot? Statistical Discrimination, Stereotypes, and Employer Decision Making. *The Annals of the American Academy of Political and Social Science, 621*, 70-93.

Phelps, E. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review, 62*(4), 659-661.

Riach, P., & Rich, J. (2002). Field Experiments of Discrimination in the Market Place. *The Economic Journal*, *112*(483), F480-F518.

Rooth, D. (2007). Implicit Discrimination in Hiring: Real World Evidence. *Social Science Research Network.* https://ssrn.com/abstract=984432.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

Vuolo, M., Uggen, C. & Lageson, S. (2018). To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis. In S. M. Gaddis (Ed.) *Audit Studies: Behind the Scenes with Theory, Method, and Nuance* (pp. 119-140). Springer.

Vuolo, M., Uggen, C., & Lageson, S. (2016). Statistical Power in Experimental Audit Studies: Cautions and Calculations for Matched Tests with Nominal Outcomes. *Sociological Methods & Research*, *45*(2), 260-303.

Wright, B., Wallace, M., Bailey, J., & Hyde, A. (2013). Religious Affiliation and Hiring Discrimination in New England: A Field Experiment. *Research in Social Stratification and Mobility*, *34*, 111-126.

Zussman, A. (2013). Ethnic Discrimination: Lessons from the Israeli Online Market for Used Cars. *The Economic Journal*, *123*(572), 433-468.

11 Appendix

Table A1 - Distinguishing English from Muslim and 'European' Names

Muslim Names		'European' Names	
Name	*Mean*	*Name*	*Mean*
Ahmed	0.97**	Antoine	1.00**
Mohammed	1.00**	Francisco	0.94**
Fatima	0.86**	Francisca	0.91**
Jamila	0.93**	Valérie	0.94**
Oliver	0.00**	Oliver	0.09**
Charlie	0.00**	Charlie	0.00**
Gary	0.00**	Gary	0.03**
Emily	0.00**	Emily	0.00**
Donna	0.00**	Donna	0.00**
Olivia	0.00**	Olivia	0.09**

Notes: n = 35. Mean scores equal the fraction of respondents who picked out that name as Muslim or European (but not English). The null hypothesis is the mean equals the average score of the other group (e.g. that the mean score for 'Ahmed' equals the average score for non-Muslim names). p values calculated with a two sided binomial test. ** denotes $p < 0.01$.

A2 - 'Lower Class' Names	
Name	Mean
Kevin	0.53
Gary	0.63
Connor	0.53
Jade	0.93
Stacey	0.63
Donna	0.67

Notes: A name's mean corresponds to the fraction of the sample who picked out that name as 'lower class'. n = 30.

A3 - Mohammed vs Gary	
Name	Mean
Mohammed	0.19
Gary	0.81
p value	0.00

Notes: A name's mean corresponds to the fraction of the sample who picked out that name as 'lower class'. The null hypothesis is that each respondent was equally likely to choose Gary and Mohammed. n = 27.

TABLE A4 - BELIEFS ABOUT TERRORISM	
Mean (%)	42.92
Standard Error	4.33
Median	40
Mode	30
Standard Deviation	30.63
Minimum	0
Maximum	98

Notes: n = 50. Respondents were asked to estimate the percentage of successful, failed or foiled terrorist attacks in the UK between 2005 and 2016 inclusive that were conducted by individuals who professed Islamic goals. The correct answer is around 2%.

TABLE A5 - DOES INTERACTION MATTER? I				
	(1)	(2)	(3)	(4)
Constant	0.43***	0.53***	0.81**	0.86**
Applicant is Muslim	-0.10***	-0.12***	-0.10***	-0.12***
Property is unshared	0.17*	0.12	-0.01	-0.11
Interaction	-0.06	-0.04	-0.06	-0.04
Covariates?	No	No	Yes	Yes

Notes: Columns (1) and (2) report coefficients from a regression of a 'yes' response on applicant religion, unshared property and their interaction. Columns (3) and (4) include the interactions of host and property covariates with the religion of the applicant. Columns (1) and (3) use the 'narrow' definition of 'yes'; columns (2) and (4) use the 'broad' definition. * denotes p < 0.05, ** denotes p < 0.01 and *** denotes p < 0.001.

TABLE A6 - DOES INTERACTION MATTER? II

	(1)	(2)	(3)	(4)
Constant	0.33***	0.44***	0.56***	0.68***
Applicant is Muslim	-0.06**	-0.10***	-0.06**	-0.10***
Live-Out Landlord	0.17***	0.17***	0.17***	0.16***
Interaction	-0.04	-0.05	-0.04	-0.05
Covariates?	No	No	Yes	Yes

Notes: 'Live-Out Landlord' is a dummy variable that equals 1 if the landlord is a live-out landlord or estate agent. Columns (1) and (2) report coefficients from a regression of a 'yes' response on applicant religion, Live-Out Landlord and their interaction. Columns (3) and (4) include the interactions of host and property covariates with the religion of the applicant. Columns (1) and (3) use the 'narrow' definition of 'yes'; columns (2) and (4) use the 'broad' definition. ** denotes $p < 0.01$ and *** denotes $p < 0.001$.

TABLE A7 - EXPLORATORY ANALYSIS

	Narrow Yes	Broad Yes
Female Applicant	-0.02	0.00
Male Landlord	-0.04	-0.02
London	-0.04	-0.02
Manchester	0.01	0.01
Birmingham	0.00	0.01
Bradford	0.01	-0.13
House Share	0.01	0.00
Flat Share	0.02	0.02
House to Rent	0.25	0.27
Flat to Rent	-0.12	-0.10
Live-In Landlord	0.04	0.01
Live-Out Landlord	-0.09	-0.07
Estate Agent	0.02	0.04
Current Flatmate	0.05	0.03
Former Flatmate	0.02	0.04

Notes: This table reports regressions of a 'yes' response on the ethnicity of the applicant, an explanatory variable and the interaction of the two. The base categories are Leeds (for cities), shared property of unknown type (for properties) and current tenants (for landlords). For the first and second columns use the 'narrow' and 'broad' definitions of 'yes' respectively. All p values exceed 0.1.