Labor market discrimination of minorities? yes, but not in job offers

Martin Bøg and Erik Kranendonk

Erasmus School of Economics

15. April 2011

Labor Market Discrimination of Minorities? Yes, but not in Job Offers*

Martin Bøg[†], Erik Kranendonk[‡]

April 15, 2011

*We gratefully acknowledge the assistance of the city of Nijmegen in providing us with the data.

[†]Corresponding Author. Erasmus University Rotterdam, Department of Economics, Room H08-01, Postbox 1738, 3000 DR Rotterdam, Netherlands. Ph.: +31 (0)104081479. Email: mbog@ese.eur.nl

[‡]Current Affiliation: Dutch Ministry of Social Affairs and Employment. Postbus 90801, 2509 LV Den Haag, Netherlands. This paper was written while the author was affiliated with Erasmus University Rotterdam. The views expressed are those of the author and do not necessarily represent the views of the Dutch Ministry of Social Affairs and Employment.

1 Introduction

Anecdotal evidence and a quick look at employment statistics suggest large differences in group outcomes as measured by the accumulation of skills, earnings and employment rates. This phenomenon is prevalent and persistent across developed economies. Field evidence and wage decomposition exercises alike support these observations (Riach and Rich 2002).

Broadly speaking there are two explanations for the large group discrepancies in labor market participation and employment rates. The first view holds that the labor market is essentially free of racial stereotyping, instead the cause can be traced back to differences in "initial conditions", that is differences in pre-market factors valued by the market(Neal and Johnson 1996). Proponents of the second view hold that ethnic stereotyping is persistent and substantial in the market. When differential treatment of identical agents is present in a market, the decisions of economic agents will be distorted. In the context of the labor market such distortions affect the incentives to apply for jobs and to invest in human capital (Coate and Loury 1993, Lundberg and Startz 1983). While differential treatment may be efficient if informational asymmetries are severe (Norman 2003), the distribution of income in such a society will be characterized by inequality. The policy implications from these two views are radically different. The former view implies that scarce resources should be directed towards correcting inequalities in pre-market conditions, whereas the second view focuses on market regulation.

The present study contributes to this debate by providing evidence from a field experiment conducted in the public sector of a major Dutch city[1]. The experiment was a policy experiment designed to test the efficacy of anonymous application procedures. The main ingredient of such a procedure is to hide group membership from recruiters at the initial stage of application, that is the stage where recruiters decide whom to interview. Although the policy evaluation is of separate interest in itself, more importantly we observe outcomes for both stages of the application process, namely first the selection of the pool of candidates invited for interview, and secondly those candidates who receive a job offer. The experiment has a classic control/treatment design. In the control group recruitment procedures are unchanged[2]. Recruiters in the control group base their interview decision

[1]The context and institutional background is described in section 3.

[2]Recruiters in the control group were aware that there was an experiment, see section 3.2.

on their assessment of a resume and an accompanying letter. In the treatment group the experimental manipulation consists of removing information from the application package that is perceived to be markers of group membership (such as name and email address. See section 3.2 for details). Recruiters who are treated make interview decisions under a (partial) veil of ignorance. By comparing the decisions made in the control and treatment group we can quantify the role of group membership in the interview decision.

The experimental manipulation used in this field experiment differs from other field experiments designed to detect differential treatment in the labor market. Two other designs have prominence in the literature: audit and correspondence studies. In audit studies the experimenter trains/instructs actors to attend job interviews. The actors are matched on observable characteristics, except the group marker. A comparison of (net-)success rates then allows for *a* measure of differential treatment. The main concerns raised in the literature regarding this research design centers around the presence of experimenter effects, and whether the two candidates used look identical (apart from group membership) in the eyes of the recruiter. The advantage of the audit design is that it allows for the possibility to observe the job offer decision. Correspondence studies[3] make a different trade-off. At the cost of focusing on the interview decision only, they gain additional experimental control. In particular correspondence studies rely on sending fictitious resumes in response to job openings. Group membership is randomly assigned to a resume and pairs of resumes are then sent in response to job openings. The measure of differential treatment is then based on differences in the (net) call back rate.

While correspondence studies retain substantial experimental control, they are prone to the critique that in equilibrium minority members will not send applications to discriminating employers. Therefore these studies tend to overestimate the extent of differential treatment present in the *market* (Heckman 1998). The present study avoids this problem. We observe all applications submitted by market actors in a specific local labor market over a 6 month period. In addition the fact that the experimental design has a classic control/treatment element allow us to rule out a number of competing explanations for the differential treatment we observe.

In our study as opposed to audit and correspondence studies the pool of applicants

[3]See e.g. Bertrand and Mullainathan (2004) for a recent field experiment utilizing this research design.

is the population of interest. Also our measure of discrimination can be interpreted as a measure of (local) market discrimination that is discrimination at the margin. Since we observe all applications in a local labor market its is reasonable to expect that market actors have internalized any differential treatment in their (costly) search decision. An additional advantage is that we have data for the second stage where job offers are made.

We find robust and statistically significant evidence for differential treatment based on group membership in call back rates. The estimates are economically meaningful. For an applicant with average characteristics this amounts to a majority applicant having to send around 6 applications to land an interview whereas her minority counterpart needs to send roughly 13 applications. These results are in line with the findings of Bertrand and Mullainathan (2004), who use a correspondence methodology, for the US of a racial gap of 50%. An interesting question remains whether these large differences translate into different offer rates? One interpretation of the differential treatment we observe is that minority workers face tougher standards. Conditional on making it to the interview stage we would thus expect the pool of minority workers to be better qualified than the workers in the majority pool. Indeed our estimates suggest that once at the interview stage the minority worker faces positive differential treatment (although this effect is not statistically significant). We do not find statistically significant evidence for differential treatment based on ethnicity in the unconditional offer rates, although our point estimates suggest a small advantage bestowed upon majority workers.

These findings are interesting for several reasons. First, we credibly estimate a large group gap in the call back rate. While we are not the first to do so, the methodology of the experiment allow us to partially side step critiques based on market sorting. Indeed since we observe all "attempts to trade" we would expect sorting to have already taken place. Second the experimental setup allow us to rule out a number of possible explanations for the differential treatment. Since we cannot condition on all information contained in the CV's and application letters a genuine concern is that the group marker is correlated with some variable that the recruiter observes and which is directly related to productivity, but this variable is not available to the econometrician[4]. The experimental setup allow us to rule out that such an omitted variable is the driver of our findings. Indeed in the treatment where

[4]Essentially explanations relying on omitted variable bias (Heckman 1998).

recruiters cannot see the group marker we find no evidence of differential treatment. Third, an advantage of giving up the experimental control of e.g. a correspondence study is that we are able to look beyond the call back stage. One of the most interesting conclusions from our study is that while there is significant differential treatment at the call back stage, group membership is not important in explaining the unconditional offer decision. Significant discrimination in the call back rates may be consistent with no discrimination at the offer stage. From this observation one cannot draw the conclusion that differential treatment at the call back stage has no economic effects. Indeed the incentive to acquire skills will be distorted. Moreover minority workers will have to search more than their majority counterparts to land an interview. This suggests important distortions between the two groups e.g. in the time between jobs. Finally, this paper contains the evaluation of a policy which aims to remove differential treatment at the call back stage. A few other papers exist (which we review in the next section) but these studies tend to suffer from selection effects, mainly because in these studies recruiters choose whether to participate, and in what role they would participate (as control or treatment). Since our study covers the full labor market for local public sector jobs, the participation decision is exogenously imposed. Also recruiters do not themselves choose whether to be in control or treatment. In addition the design minimizes applicant selection effects. It is not announced which departments are assigned to control or treatment, although the experiment itself is announced. We find that no differential treatment can be detected when operating under the policy, as such the primary aim of the policy: to remove any differential treatment, is successful.

The remainder of this paper is organized as follows. Section 2 discusses related evidence in more detail. Section 3 describes the experimental setting and the design of the experiment. In section 4 we describe and analyze the data. Section 5 discusses related evidence from a follow up experiment. Finally section 6 concludes.

2 Related Evidence

There is a relative large literature exploring group differences in labor market outcomes. The literature has been particularly interested in differences based on gender and "race" (Altonji and Blank 1999). Two research methodologies are used to identify and quantify group differences: observational and experimental studies.

Observational studies typically utilize large administrative data sources to quantify groups differences in outcomes[5]. For US data large and persistent differences in earnings, the so called racial wage gap, have been found. While few disputes these differences, the literature differs on the cause. One strand of the literature attribute these differences to group differences in pre-market productivity related factors (Neal and Johnson 1996)[6]; the other strand argues that differences can be traced back to discriminating recruiters. The latter studies frequently make use of experimental methods to assess the extent of discriminatory practices at one or more layers of the recruitment process.

Experimental studies fall in two categories: audit and correspondence studies. In audit studies the experimenter trains two actors from different ethnic groups. Actors are chosen to be similar on all observable characteristics to allow for a "ceteris paribus" comparison. This methodology has revealed substantial net racial discrimination (Riach and Rich 2002). A potential problem with audit studies, apart from the significant costs needed to get a sufficiently large data set, is that the actors are informed about the nature of the study. Thus although employers are not aware that an experiment is ongoing, the measure of discrimination could be affected by experimenter effects, e.g. actors may consciously or sub-consciously try to confirm the experimenters conviction. An additional issue is what is actually measured. The discrimination measure is typically the average discrimination (averaged over employers) per audit pair. An overall measure of discrimination is then computed by averaging over pairs. The sample average may not be the population average of interest.

Correspondence studies avoid experimenter effects by relying on fictitious applicants[7]. Similar resumes are constructed and the experimenter manipulates the racial markers of

[5]A common approach to quantify group differences, is a racial dummy in an Oaxaca wage decomposition. However other approaches are used in this tradition. E.g. in the context of the Swedish labor market Eriksson and Lagerström (2007) uses a large internet-based database of resumes of the Swedish Public Employment Office. Marrying this data with administrative data they find that minority membership is correlated with a fewer number of approaches by employers and a fewer number of offers.

[6]Neal and Johnson (1996) cleverly circumvent the endogeneity problem that all observational studies must "solve", and instead ask how much of the racial gap can be explained by pre-market initial conditions only, as measured by AFQT tests.

[7]This method was pioneered by UK sociologists to measure the extent of labor market discrimination.

the resumes. Bertrand and Mullainathan (2004) uses this methodology in their study of labor market discrimination in Chicago and Boston. In response to help wanted ads they send fictitious resumes. Since group membership is typically not included in an application they instead use "stereotypical" African-American and White names in the experimental manipulation. They find large and statistically significant differences in call back rates. The African-American marker results in a 50% lower callback rate compared to the White marker. Our findings for call back rates match these findings. In a European setting Carlsson and Rooth (2007) used correspondence testing to investigate discrimination in the Swedish labor market. They use a similar manipulation to that of Bertrand and Mullainathan (2004): they randomly assigned native and Middle-Eastern "sounding" names to fictitious resumes and sent them in response to help wanted ads in Gothenburg and Stockholm. They find a 10% lower callback rate for the minority group, matching the findings of Bertrand and Mullainathan (2004) of a call back rate which is 50% lower for minorities[8].

A drawback of correspondence studies is that market transactions are never consummated, indeed since there are no real applicants the measure of differential treatment is the (net) difference in call back rates. The question of interest may not be whether the interview decision is distorted, but rather whether as a result of facing a lower chance of landing an interview minority candidates also have a lower chance of receiving job offers. By construction a correspondence study cannot address this issue. Our study speaks to this question. Indeed we find statistically significant large effects of race on the call back rate, but we only find weak evidence of differential treatment based on group membership in unconditional offers. This is consistent with a selection procedure with tougher initial standards for minorities and reverse discrimination at the interview stage (conditional offer rates).

This paper also contains an evaluation of a policy specifically aimed at addressing unequal treatment in job applications. The policy seeks to achieve this goal by concealing the group

[8]Carlsson and Rooth (2008) includes three groups: natives, native workers with Middle-Eastern sounding names and Middle-Eastern immigrant workers. They find a 17% lower callback rate for native applicants with a Middle-Eastern sounding name and a 21% lower callback rate for the immigrated applicants compared with native applicants. They conclude that the foreign name of an applicant explains 77% of the discrimination by employers.

membership at the initial stage of application.

Goldin and Rouse (2000) uses variation in the use of screens in auditioning for large symphony orchestras in the US to test for the efficacy anonymous application procedures. They are interested in the question whether the chances of females to progress increase when screens are used. They find that the use of screens increase a female's chance of progressing by 50%. In addition the use of screens in the preliminary round also increases the likelihood that a female applicant wins the final by 30% (the final round is typically not blind).

Closer to our setting Åslund and Skans (2007) perform a field experiment on the use of anonymous application procedures in the public sector of 3 districts in the city of Gothenburg, Sweden. In the treatment group both ethnicity and gender was hidden from recruiters. They find evidence of differential treatment of non-western immigrant and female applicants at both the interview and the offer stage. Non-western immigrants have a 8.9% lower probability to be invited for an interview and a 2.1% lower probability to be hired. For women the probability to be invited for an interview is 6% lower and the probability to be hired is 3.8% lower. The anonymous job application procedure increases both the probability of immigrants and women to be invited for an interview by approximately 8%. The hiring probability for females increase by 7%, but no such effect is found for immigrants. The experimental procedures of Åslund and Skans (2007) differ from ours in several important aspects. The three districts in their sample are recruited among all city districts. In addition districts are allowed to self-select into control (comparison) or treatment. This affects interpretation of their estimates. Job ads for vacancies in the treatment group indicate that an experiment is in progress and in addition to submitting a normal application applicants must fill in a special application form which does not contain information on gender and ethnicity. Recruiters in the treatment group make their decision based on the special form only, and only after a decision has been made do they receive the full application package. A complication arises here since recruiters in comparison and treatment groups do not have access to the same information over and above ethnicity and gender variables. In addition the announcement that a vacancy was in the treatment group raises concerns about possible applicant selection effects. Our experimental design avoids these potential problems.

3 Experimental Design

3.1 Setting and Institutional Background

The Netherlands is a small open economy with approximately 16 million inhabitants. The Netherlands is and has been ethnically mixed for an extended period of the modern era. In addition in the late post-war period there was a significant inflow of labor migration to the Netherlands. The Netherlands remains one of the most ethnically mixed European countries. Roughly 80% of the Dutch population is ethnically Dutch[9]. The largest non-western immigrant groups are composed of Dutch Indonesians[10], Turks, Surinamese and Moroccans (each group is approximate 2% of the total population).

In spite of a long history of integration survey evidence suggests that discrimination and stereotyping is not uncommon, both in the population at large and among recruiters. Surveys of the attitudes of the Dutch population towards minorities (Scheepers, Eisinga, and Linssen 1994, Verberk, Scheepers, and Felling 2002) find that 20% have strong negative stereotypes, and 48% have "subtle" negative stereotypes about immigrants. Unsurprisingly these negative stereotypes about minorities can also be found among employers. Kruisbergen and Veld (2002) survey employer attitudes towards young immigrant workers. They find that 6% of employers would never fill a vacancy with an immigrant worker, 18% would only hire an immigrant worker if no native worker applies.

More direct evidence on differential treatment comes from correspondence and audit studies. Bovenkerk, Gras, Ramsoedh, Dankoor, and Havelaar (1995) used audit tests to assess the extent of discrimination against males of Moroccan and Surinamese origin. For both groups significant levels of net discrimination in the call back rate was identified (bearing in mind that these are not direct measures of market discrimination). Their study show a minimum net rate of discrimination during the first phase of the job assignment process of 32% for low educated male applicants of Moroccan origin and a rate of 40% for the same group of applicants from Surinamese origin. Using correspondence tests the authors find that highly educated male applicants of Surinamese face a minimum net rate of discrimination of 18%.

[9]For comparison the UK has approximately 90% white majority.

[10]This group is considered to be western immigrants in official statistics.

The findings for the labor market stand in contrast to Article 1 of the Dutch Constitution which asserts equal treatment of its citizens. Perhaps for this reason several Dutch cities have been willing to experiment with job assignment procedures which attempt to ensure equal treatment of equally qualified candidates. Cities including Amsterdam, Nijmegen and The Hague have experimented with versions of anonymous applications procedures (AAP).

The Dutch city where this field experiment was conducted is in the top 10 of Dutch cities as measured by population size. The composition of the city population is close to the Dutch average. In 2007 8.3% of the population were of non-western origin, compared with a national average of 10.3%. The largest non-western immigrant groups are Turks (24.3%), Antillians (10.7%), Surinames (10.7%), Former-Yugoslavia (8.6%) and Moroccans (6.9%)[11]. In 2007 the unemployment rate in the council was low (5.9% for 2006/2008 compared with 7.1% for 2005/2007)[12]. This is in line with the general trend in unemployment rates for the Netherlands in that period[13].

3.2 Experimental Procedures

In July 2006 the city council and the mayor approved the experiment with anonymous application procedures. The experiment ran for a 6 month period from August 1st 2006 until February 1st 2007. During this period all new vacancies were included in the experimental sample either as control or treatment. In the taxonomy of Harrison and List (2004) this experiment is a framed field experiment. The recruiters who participate in the experiment take on the roles that they normally have. If assigned to treatment not all information that is normally available to recruiters is present, but the task itself, that of scrutinizing application letters and resumes and deciding who to interview is familiar to recruiters. The second stage of the applicant screening, the interview, is not affected by the experiment. In particular recruiters from departments that are assigned to treatment also meet face-to-face with candidates at the interview stage. For an overview of the application procedure

[11]Source: Dutch Statistical Agency (CBS). `http://www.cbs.nl`.

[12]Source: `http:\www.cbs.nl`.

[13]Unemployment rates for the Netherlands as a whole was 5.5% (2006), 4.5% (2007) and 3.8% (2008). In 2007 and 2008 the unemployment rate was considered to be below the natural rate of unemployment by the CBS.

and the experimental intervention see figure 1.

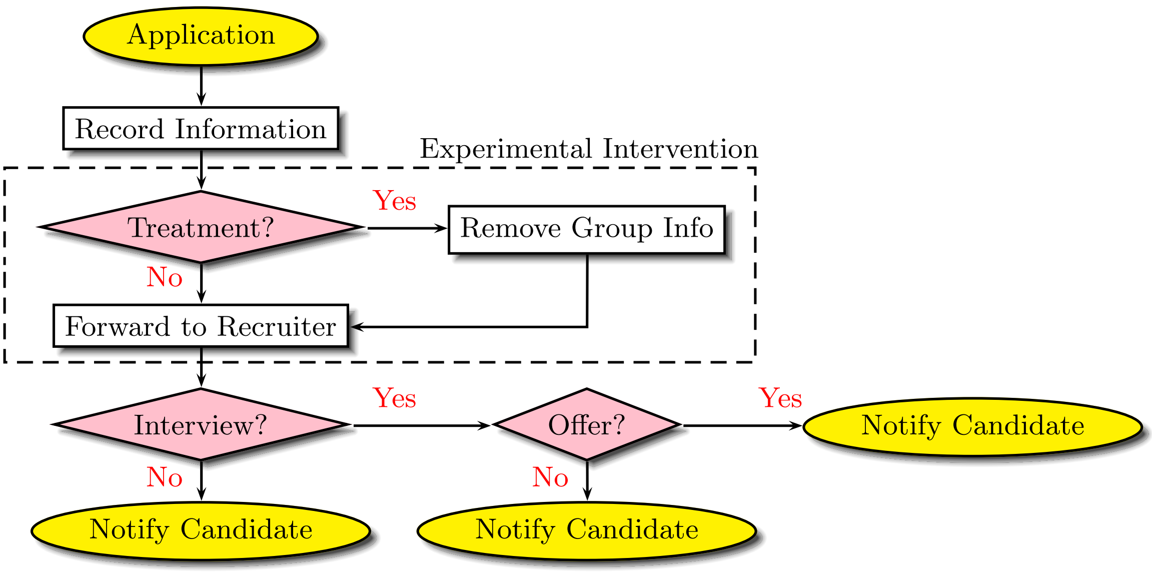


Figure 1: Job Application Procedure. Experimental Intervention in dashed box.

Experimental Assignment The local public administration consists of 7 departments. Each department is assigned to either control or treatment for the duration of the experiment[14]. It is important to stress that the assignment to treatment or control was not random, rather it was decided administratively. The decision was exogenous to participating departments, but as indicated below the joint characteristics of the two groups were matched. The experiment designers made the decision whether a particular department should be assigned to control or treatment. The only objective for assignment to either treatment or control was to get two groups that were comparable in size in terms of total employment and in fraction of non-western immigrants employed. After experimental assignment the treatment (control) group consisted of a group of departments with 1055 (909) employees, and an immigrant stock of 6% (9%). We stress at this point that it is crucial for the interpretation of our results that departments cannot themselves influence

[14]The 7 different departments were: "Stadsbedrijven" (Public enterprizes), "Grondgebied" (Territory), "Wijk en Stad" (District and City), "Inwoners" (Resident services), "Bestuursstaf" (Groups Staff), "Concernstaf" (Administrative Staff), "Brandweer" (Fire department). The first 3 departments were assigned to the treatment group, while the remaining 4 departments were assigned to the control group.

whether they are in treatment or control[15].

Posting of Vacancies and Informational Setting During the experimental period vacancies were posted through normal job channels. Job advertisements did not mention that an experiment was in progress. However the experiment did receive press coverage, both in local radio, tv and newspapers[16]. Therefore we cannot exclude that applicants were aware that an experiment was in progress. Importantly it was not made public whether a department had been assigned to treatment or control. The experimental design allows us to exclude applicant selection effects as a possible driver of our results. Such effects would be present if an applicant's decision on whether to apply for a position depended on her knowledge of whether the vacancy was in control or treatment. We cannot rule out that the public announcement of an experiment has affected the sample on the extensive margin. That is the public announcement may crowd in applicants who would not otherwise apply, perhaps due to the belief that differential treatment is prevalent in the application procedure.

Processing of Applications An application consists of a letter and a resume. During the experimental period all applications for vacancies goes directly to a working unit who processes the applications and encodes information. The task of the unit is essentially twofold: (1) Note down information about applicants and vacancies, including classifying applicants into minority status or not (see below), (2) Process applications for vacancies in the treatment group.

The unit collects the information about the applicants. This information consists of gender, age, educational level, and whether the candidate is an external candidate or not. Importantly the perceived minority status of the candidate has to be determined. Whether the candidate has minority status or not is typically not revealed directly in the application. Furthermore the mapping between having official minority status and the perception of such is not one-to-one[17]. Instead for the purpose of the experiment an applicant has minority

[15] To be precise our estimates would still be meaningful but only as e.g. an intention-to-treat population estimator.

[16] For instance the experiment was mentioned in an article in the national newspaper *NRC-Handelsblad* on 25th January 2006 `http://archief.nrc.nl/?modus=l&text=anoniem+solliciteren+nijmegen&hit=6&set=2&check=Y`.

[17] Applications do not typically list e.g. nationality or place of birth which is typically used by adminis-

status if s/he has a non-western foreign "sounding" last name. Thus e.g. applicants with a German, French or English last name are not classified as having minority status. This classification differs from the classification applied both by Dutch Authorities and the Dutch Statistical Agency (CBS)[18]. Applicants are not asked to self-identity rather a team of 3 independent testers determines whether the name is foreign "sounding" in the sense described above.

In addition the two outcome variables are collected. That is whether or not the candidate is invited to attend a job interview, and if so whether a job offer is made following the interview.

The team also collects summary information about the vacancy. The following information is collected: job title, announced educational requirements, whether vacancy is permanent or temporary, date of vacancy and whether the vacancy is in control or treatment[19].

Whether applications are manipulated further before recruiters have a chance to look at the applications depends on whether the vacancy is in the treatment or the control group. If the application is in the control group the application is forwarded "as is" to the relevant recruiters. On the other hand if the application is in the treatment group then the following steps are undertaken before the application is forwarded to recruiters. Information on name, email address, birthplace, country or origin and nationality of the applicant is "tippexed" (whitened out). Only after this manipulation has been completed is the application forwarded to the relevant recruiters.

The next step in the application procedure for a vacancy is the ranking of the set of submitted applications. This step is identical for vacancies in the control and treatment group, except for the fact that some information has been whitened out for applicants to vacancies in the treatment group. Recruiters make a decision on whom to interview (typically more than one candidate is called for an interview, see table 2). After compiling their ranking of candidates, recruiters then communicate their decision to the experimental unit who contacts candidates. Once at the interview stage there is no interaction between

trative bodies as the basis for ethnic classification.

[18]The classification is similar to the one used in the correspondence study of Bertrand and Mullainathan (2004).

[19]Information about which department had the vacancy was not recorded.

the experimental work unit and the recruiters apart from taking note of the outcome of the interviews. To be clear once at the interview stage there is no experimental intervention taking place; the experimental manipulation is only indirect through the selection of candidates.

3.3 Comments on Experimental Setup

At first look the ideal experimental setup randomly assigns each vacancy to either control or treatment. This would allow the analyst to average out vacancy specific effects. For practical reason such a setup was not feasible. Instead in our analysis we shall take account of vacancy specific effects. A more fundamental reservation against the "ideal" experimental setup is the following. Suppose that recruiters harbor belief-based negative stereotypes against the minority, while in reality a minority applicant is equally qualified viz-a-viz her majority counterpart. Such beliefs may lead the recruiter to abstain from further scrutinizing minority applicants. Since under normal recruitment procedures the recruiter seldom encounters a qualified minority applicant, she can harbor these stereotypes. However by being exposed to minority applicants (as she would very likely be under an anonymous application procedure) the recruiter may gain new information about applicant qualifications and update her beliefs accordingly. If recruiters adapt their hiring procedures fairly fast to this new information the proposed "ideal" experimental procedure which randomizes within-departmental vacancies could fail to detect discrimination although present and accordingly would also conclude that the anonymous application procedures policies is unnecessary.

3.4 Potential Confounds

Applicant Margin Although the city council itself did not undertake considerable specific efforts to make the experiment known to all potential applicants, the fact that such a large scale experiment was about to be put into place naturally received attention in the local press. Although we do not have access to data that would allow us to quantify an effect of the "announcement" we cannot exclude that applicants react to this information. Applicants on the margin may have been crowded either in or out of the experimental sample. A majority marginal candidate (marginal to applying that is) who expects to benefit from standard recruitment practices may choose to abstain from applying when s/he is aware that there is approximately 50% chance that the vacancy is in the treatment group. Symmetrically the announcement may crowd in applications on the margin from minority workers who expect equal treatment if the vacancy is in the treatment group.

A related concern is that our estimates may be affected by applicant selection effects. Such an effect would be present if applicants sort into vacancies depending on whether they are in the treatment or the control group. Our estimates are not affected by this effect. First, job adverts did not mention that there was an experiment in place[20]. Second, and more importantly it was never made public which departments were assigned to treatment and which to control. Therefore by design we can exclude applicant selection effects.

Recruiter Margin Recruiters whether assigned to treatment or control were aware that there was an experiment[21].

Two well known experimental effects may thus affect the interpretation of our estimates. A well known experimental effect is the John Henry effect. This experimental effect suggest that subjects assigned to the control group would work harder at behaving like the treatment group. This would tend to bias our measure of differential treatment downwards[22].

A second potential experimental effect is the Hawthorne effect. This effect suggest that any change in the environment of agents improve their behavior (in the direction of the principals goals). The presence of this effect would suggest that our measure of differential treatment would be biased downwards. Indeed in this experiment the mere fact that the principal clearly signals something about her objective function would suggest that recruiters work harder at applying standards which do not depend on group membership. If the Hawthorne effect is present in our study we would therefore expect it to bias our estimates downwards. It may be argued that the treatment group received more attention than the control group, and that the effect of an anonymous application procedure

[20]As opposed to other experimental studies in this class applicants were not asked to fill out additional forms declaring minority status, etc.

[21]We return to this issue in section 5.

[22]Post-experimental surveys and interviews with recruiters were also conducted. According to these interviews recruiters believed that they did not treat applications differently depending on the ethnicity of the applicant.

would vanish over time. It may be that "tippexing" the most salient group membership information is not sufficient to hide the ethnic identity of the applicant. E.g. recruiters may use clues from the letter style or the residence of the applicant to deduce her ethnic identity. It could affect the long run efficacy of anonymous application procedures as it was implemented in the present study. If the Hawthorne effect is present, we may surmise that after anonymous application procedures are implemented across the board, and experimental "attention" vanishes, then clues would again become salient. This concern, does not affect, however, the extent of differential treatment as measured by the experiment.

4 Empirical Analysis

This section contains our main empirical results. We first provide a summary overlook of our sample and then we proceed to estimation. Finally we explore whether the experimental design allow us to shed any light on which type of differential treatment is at work in our data.

4.1 Data

The data generating process for the data used in this study is as follows. Each application sent for a vacancy generates a data point. For each data point the information disclosed to us fall in three categories[23]: (1) applicant covariates, (2) vacancy covariates and (3) outcome variables. The applicant covariates includes: age category (broken down in three categories: below 30 years of age, between 30 and 45 years of age, and above 45 years of age), educational attainment of candidate (grouped into low, middle and high), gender of the applicant, whether applicant is an internal or external candidate and group membership of the candidate (based on the method of classification outlined in section 3.2). The vacancy covariates includes: job title, posted educational requirement for the vacancy, whether the vacancy is permanent or temporary, date that vacancy was posted and whether the vacancy is assigned to treatment or control. Finally we observe two (binary) outcome variables, namely whether the candidate is invited for interview, and whether the candidate received a job offer. Our data set contains a total of 1200 applicants spread over 37 vacancies.

[23]Before we got access to the data it was anonymized and screened for confidentiality.

4.2 Descriptives

4.2.1 Vacancies

Table 1 gives a first look at the data, in terms of the number of vacancies posted in treatment and control group and the number of applicants. The table is also broken down on the educational requirements for the vacancies.

During the experimental period 17 vacancies were posted in the treatment group, while in the control group 20 vacancies were posted. When breaking these numbers down by educational level required the main difference is that the control group has a slightly higher number of vacancies for the highest educational level. These differences are not statistically significant[24]. Turning to the number of applicants the sample contains a total of 1200 applicants. 663 applicants applied for the vacancies in the treatment group, while 537 applicants applied for the vacancies in the control group. When these differences are broken down by educational requirements we do see some heterogeneity across treatment and control in terms of number of applicants and these differences are statistically significant[25]. For some posted vacancies several positions are available. Although we do not know the exact number of available positions per vacancy the number of offers made provide a lower bound on the number of positions (for all vacancies at least one job offer was made). For the posted vacancies in the treatment group a total of 20 job offers were made, while for the control group a total of 36 job offers were made. These differences highlight the importance of controlling for vacancy specific effects when we proceed to estimation.

Required	**Number of Vacancies**			**Number of Applicants**		
Education	Treatment	Control	Total	Treatment	Control	Total
Low	10	11	21	479	356	835
Medium	5	5	10	155	80	235
High	2	4	6	29	101	130
Total	17	20	37	663	537	1200

Table 1: Vacancies and Applicants by Required Education Level

Table 2 provides a closer look at the variation within vacancies across control and treatment

[24] $p = .814$ using Fisher Exact test.

[25] $\chi^2 = 69.47$, $p = 0.000$.

broken down by educational level required. It is worth noticing that there is substantial variation in the number of applicants per vacancy (partially explained by the earlier observation of several openings per posted vacancy). Vacancies at the lower educational levels tend to receive more applicants. None of the differences, between control/treatment and within control/treatment are statistically significant (Mann-Whitney test).

Required	**Treatment**			**Control**		
Education	Applicants	Invited	Offered	Applicants	Invited	Offered
Low	47.9	.173	.098	35.5	.185	.112
	(37.262)	(.163)	(.152)	(26.082)	(.127)	(.108)
Medium	31	.175	.072	16	.223	.093
	(39.956)	(.106)	(.049)	(5.788)	(.143)	(.035)
High	14.5	.338	.101	25.25	.251	.071
	(4.950)	(.164)	(.014)	(19.414)	(.206)	(.055)
Total	39	.193	.091	28.211	.209	.098
	(36.418)	(.150)	(.117)	(21.984)	(.143)	(.083)

Notes: Means reported for vacancies with more than 1 applicant. Invitation and Offer fractions are means of vacancy specific chances. Let j be a vacancy, n_j be the number of applicants for position j and m_j the number of invited applicants. Then $\mu = \sum_{j=1}^{J}(m_j/n_j)/J$. Standard deviations in parentheses.

Table 2: Vacancy Heterogeneity by Required Education Level

4.2.2 Applicant Characteristics

We now turn to describing applicant characteristics. Table 3 presents descriptive applicant statistics for the treatment and control group. The sample is broken down by the listed educational requirement for the vacancy. In the control group 19% of applicants are minority workers, whereas the 15% of the applicants in the treatment group are minority workers. This difference is statistically significant (Mann-Whitney ranksum test, $p = .069$). Broken down by required education level, the only significant difference is for low education vacancies, where a significantly higher fraction of minority workers applied for vacancies in the control group (22% vs. 17%, $p = .079$). Turning to gender the fraction of female applicants are not significantly different for the total sample. In the control group 57% of applicants are females and 60% in the treatment group. For the low education level the fraction of female applicants are significant larger in the treatment group (76% vs. 69%, $p = .025$). The modal age for both the treatment and control fall in the age category 30-45 years of age (49% and 40% for the control and treatment group respectively). The age composition of applicants differs significantly across treatment and control (Pearson test, $p = .003$). Conditioning on required education these differences are driven by the low and middle education requirements ($p = .093$ and $p = .000$ respectively). Finally the table also provides summary information about the level of education of candidates. The modal education is "high" in the control group (52%) while middle educational level is modal in the treatment group (55%). It is worth noting that for the low and middle educational requirement vacancies applicants tend to have better qualifications than required for the vacancy, e.g. for the low education vacancies the modal educational level is middle (54% vs. 64% for control and treatment respectively). Likewise for the middle educational requirement vacancies the modal educational level is high (80% and 54% for control and treatment respectively). The educational composition of candidates differ significantly between groups ($p = .000$). When we condition the sample on the required educational level the differences across groups remain significant ($p = .000$, $p = .000$, $p = .067$ for low, middle and high respectively).

Table 4 shows a different break down of the data. In particular the table conditions on group membership and whether in treatment/control.

Compared to the majority group there is a larger fraction of minority applicants who are females. Within groups but between treatment/control there is no significant differences for either of the groups. Turning to the age distribution, there is no difference between treatment and control for the minority group. For the majority on the other hand there is a significant difference ($p = .018$). These differences are mainly driven by differences in the age composition of candidates for vacancies in the middle education category. Within this vacancy category the treatment group is significantly more senior ($p = .000$)[26].

The columns invited and offer gives the percentage of candidates invited for interview and offered a job broken down on group membership and treatment/control. The fraction of minority candidates invited does not differ substantially between treatment and control; in both the control and treatment group roughly 9% of the applicants were invited. On

[26]Table with breakdown per vacancy category not shown. Available upon request.

Required Education		**Majority**	**Female**	**Age** <30	30-45	45-65	**Education** Low	Medium	High	**N**
Low	Control	.778*	.688**	.25	.486	.242*	.0955	.542	.331***	356
		(.416)	(.464)	(.434)	(.501)	(.429)	(.294)	(.499)	(.471)	
	Treatment	.827	.758	.273	.401	.299	.148	.637	.2	479
		(.379)	(.429)	(.446)	(.491)	(.458)	(.356)	(.481)	(.401)	
Middle	Control	.813**	.3	.2	.588	.2***	.0125	.163	.8***	80
		(.393)	(.461)	(.403)	(.495)	(.403)	(.112)	(.371)	(.403)	
	Treatment	.916	.206	.0839	.361	.523	.0645	.374	.542	155
		(.278)	(.406)	(.278)	(.482)	(.501)	(.246)	(.485)	(.5)	
High	Control	.901	.366	.208	.406	.347	0	.0297	.96**	101
		(.3)	(.484)	(.408)	(.494)	(.478)	(0)	(.171)	(.196)	
	Treatment	.793	.241	.138	.586	.276	0	.138	.862	29
		(.412)	(.435)	(.351)	(.501)	(.455)	(0)	(.351)	(.351)	
Total	Control	.806*	.57	.235	.486	.255***	.0652	.389	.52***	537
		(.396)	(.496)	(.424)	(.5)	(.436)	(.247)	(.488)	(.5)	
	Treatment	.846	.606	.223	.4	.35	.122	.554	.309	663
		(.361)	(.489)	(.417)	(.49)	(.477)	(.328)	(.498)	(.463)	

Notes: Variable means. Standard deviations in parenthesis. Mann-Whitney ranksum means test for majority and gender variables. Pearson's test for age and education categories. *** p <0.01, ** p <0.05, * p <0.1. 31 observations has missing age category. 24 observation has missing education category.

Table 3: Applicant Descriptive Statistics by Required Education Level (Means). Full Sample.

the other hand for majority applicants treatment has a significant effect on the fraction invited. In the control group 16% of majority applicants are invited, while in the treatment group only 10% are invited for an interview ($p = .004$). The same differences between the two groups are visible when looking at the fraction of applicants who receive a job offer. In the minority group 4% receive an offer in control, to be compared with 3% for those in treatment (not statistically significant). For majority applicants the fraction offered a job drops from 7% to 3% when in treatment (statistically significant, $p = .002$). This suggests that treatment (if effective) works by lowering the chances for majority workers but not for minority workers. However since the treatment group has more applicants, the overall chance of invitation has decreased, but this does not take into account that there was on average more applicants for vacancies in the treatment group (see table 2). We return to this point later.

					Age			Education			
Group		**Invited**	**Offer**	**Female**	<30	30-45	45-65	Low	Medium	High	**N**
Minority	Control	.0865	.0385	.692	.394	.481	.0962	.0769	.462	.404**	104
		(.283)	(.193)	(.464)	(.491)	(.502)	(.296)	(.268)	(.501)	(.493)	
	Treatment	.0882	.0294	.686	.353	.422	.196	.0784	.637	.284	106
		(.285)	(.17)	(.466)	(.48)	(.496)	(.399)	(.27)	(.483)	(.453)	
Majority	Control	.157***	.0739***	.54	.196	.487	.293**	.0624	.372	.547***	433
		(.364)	(.262)	(.499)	(.398)	(.5)	(.456)	(.242)	(.484)	(.498)	
	Treatment	.0963	.0303	.592	.2	.396	.378	.13	.538	.314	561
		(.295)	(.172)	(.492)	(.4)	(.489)	(.485)	(.337)	(.499)	(.464)	

Notes: Variable means. Standard deviations in parenthesis. Mann-Whitney ranksum means test for invited, offer and gender variables. Pearson's test for age and education categories. *** p <0.01, ** p <0.05, * p <0.1. 31 observations has missing age category. 24 observation has missing education category.

Table 4: Applicant Descriptive Statistics by Group Membership (Means). Full Sample.

Table 5 gives summary statistics for the second stage of the application procedure, i.e. those invited for an interview. 77 candidates were invited for interview in the control group and 63 in the treatment group. The sample contains 18 minority and 122 majority candidates. Conditioning on group membership the difference in distribution of characteristics between candidates selected for treatment and control groups is small. The age distributions do not differ significantly between treatment and control for either group. Turning to education for majority candidates there is a discernible difference ($p = 0.067$) between treatment and control. This is driven by vacancies in the middle education requirement range, where the control group candidates have a higher level of education compared to candidates in the treatment group ($p = 0.046$, table not shown). The fraction of female candidates invited for interview in the minority group is 67% in the control and 44% in treatment. The fraction for majority candidates are 56% and 63% in the control and treatment group (not significant).

Finally we look at the (conditional) offer rate. Around 44% of minority candidates in the control group received an offer, while 33% of their treatment counterparts received an offer (not significant). 47% of majority candidates in the control and 32% in the treatment group received an offer (significant, $p = 0.0825$).

Group		**Offer**	**Female**	Age <30	Age 30-45	Age 45-65	Education Low	Education Medium	Education High	**N**
Minority	Control	.444	.667	.333	.556	.111	0	.333	.556	9
		(.527)	(.5)	(.5)	(.527)	(.333)	(0)	(.5)	(.527)	
	Treatment	.333	.444	.444	.444	.111	0	.556	.444	9
		(.5)	(.527)	(.527)	(.527)	(.333)	(0)	(.527)	(.527)	
Majority	Control	.471*	.559	.206	.529	.25	.0441	.368	.574**	68
		(.503)	(.5)	(.407)	(.503)	(.436)	(.207)	(.486)	(.498)	
	Treatment	.315	.63	.167	.648	.185	.0741	.574	.352	54
		(.469)	(.487)	(.376)	(.482)	(.392)	(.264)	(.499)	(.482)	

Notes: Variable means. Standard deviations in parenthesis. Mann-Whitney ranksum means test for offer and gender variables. Pearson's test for age and education categories (Fisher's exact test in case of small number of observations). *** p <0.01, ** p <0.05, * p <0.1. 1 observation has missing age category.

Table 5: Applicant Descriptive Statistics by Group Membership. Interview Sample.

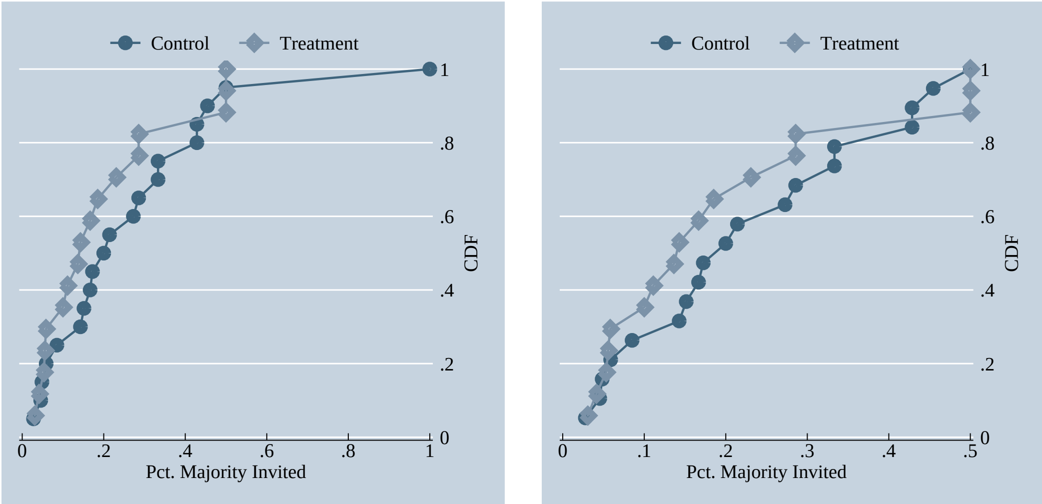
4.2.3 The Effect of Treatment: Empirical CDFs

In appendix A we introduce a simple model for the decision to interview. The model has a number of predictions about recruiter behavior in the control and treatment condition. In particular when an employer has group based beliefs about the match quality she should set different group standards. If the employer has positive stereotypes about one group relative to the other group, then she optimally sets lower standards for members of that group.

The goal of an anonymous application policy is to make the interview decision independent of group membership (all other characteristics being equal). It seeks to achieve this by masking the group identity of candidates. When AAP is effective the employer uses a single standard by which she judges all candidates. This single standard, is between the two group based standards that the recruiter sets when she can condition on group membership: $s_{maj} < \bar{s} < s_{min}$. The model therefore predicts a particular pattern in response to treatment viz-a-viz control. We now briefly investigate these prediction by looking at the empirical CDF's for the fraction of invitees of both groups and the composition of the pool of interviewees.

Figure 2 shows the empirical CDF for the chance of a majority candidate to be invited for interview. The unit of observation is a vacancy and for each vacancy the fraction of majority candidates invited is the measure used to rank the observation. The plotted distributions do not control for any heterogeneity across treatment and control, e.g. in vacancy and candidate specific covariates. The figure contains two panels. Each panel contains two graphs, one CDF for the control and one for the treatment group. In the left panel all vacancies are used in computing the CDF. In the right panel we have excluded those observations where all majority candidates were invited. The right hand panel may be more informative since for vacancies where all majority candidates are invited we cannot identify the group specific standard used.

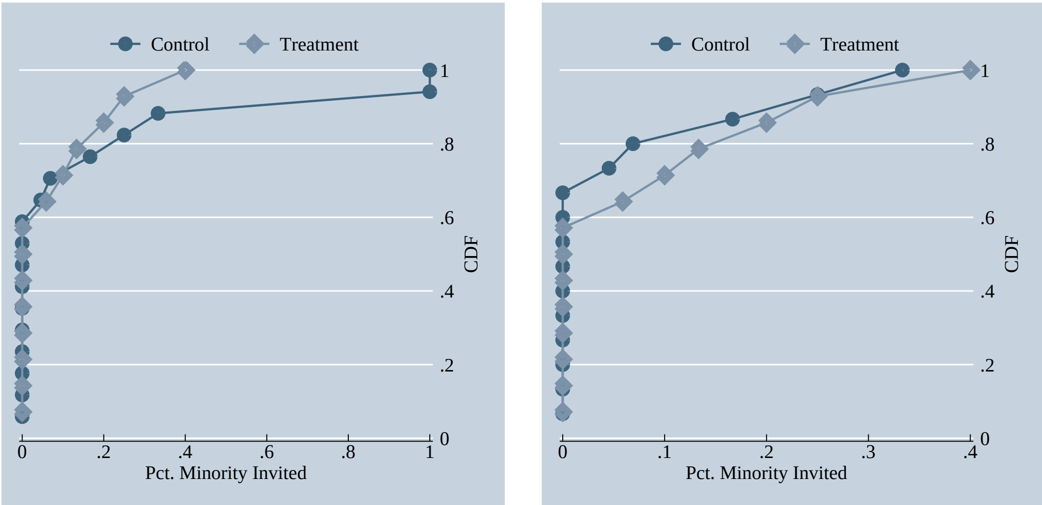
The empirical CDF cannot be ordered (e.g. by FOSD). However there is evidence of a treatment effect. In particular by and large when under treatment the chances of majority candidates to be invited for interview are lower than those of the control group. This is compatible with tougher standards for majority candidates when under treatment.



Left Panel: Empirical CDF of majority candidates invited for interview. Right Panel: Empirical CDF of majority candidates invited for interview excluding vacancies where all majority candidates invited.

Figure 2: Empirical CDF: Majority Invites

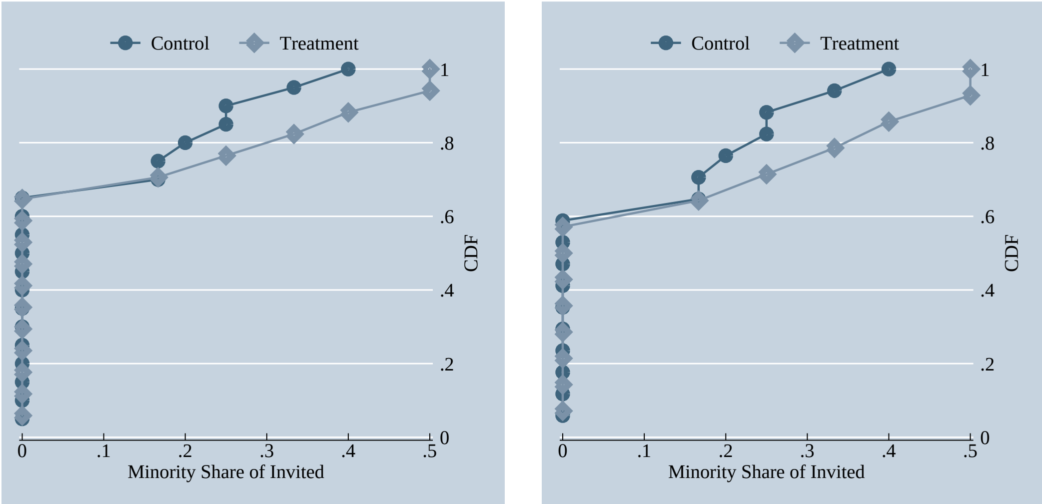
Figure 3 is similar to figure 2 except that it looks at minority candidates. As with the previous figure the right hand panel excludes vacancies where only minority candidates were invited. The model prediction is that the chances of minority candidates improves when under treatment, since the recruiter must set a common standard for all candidates. The right hand panel distributions is in accordance with this prediction. Notice also that for about 60% of vacancies no minority candidates are invited. Again since we do not condition on the observables of candidates and vacancies, the graphs are at best suggestive of an effect of treatment.



Left Panel: Empirical CDF of minority candidates invited for interview. Right Panel: Empirical CDF of minority candidates invited for interview excluding vacancies where all minority candidates invited.

Figure 3: Empirical CDF: Minority Invites

Figure 4 looks at the distribution of the vacancy specific compositions of the interview pool. The figure therefore to some extent summarizes the information from the two previous figures. The left hand panel are based on all vacancies, and the right hand panel is based only on vacancies where both minority and majority candidates *applied* for the position. Both panels show the minority share of invites in the interview pool. The panels tell a similar story. Treatment seems to shift the distribution to pools with a larger minority share of invites. This is compatible with the evidence of the two previous figures.



Left Panel: Empirical CDF of fraction of minority candidates in interview pool. Right Panel: Empirical CDF of fraction of minority candidates in interview pool. Only vacancies with both minority and majority applicants included.

Figure 4: Empirical CDF: Minority Share of Invites

4.3 Estimands and Identification

We are primarily interested in the following set questions: (1) Is there evidence of differential treatment in the interview decision, and if so is the size of the effect economically relevant? (2) If the answer to (1) is in the affirmative, is the anonymous applicant procedure successful in eliminating differential treatment? (3) Is there any evidence of differential treatment in the offer decision? This last question is of considerable interest given the current state of the literature. As outlined in the literature section there is a considerable number correspondence studies that find evidence of differential treatment in the decision to interview. Because these studies typically rely on fictitious candidates they can (by construction) only provide evidence on the interview decision and not on the job offer decision.

Before proceeding to testing for potential treatment effects it is useful to define the estimands and discuss identification. For our setting it is natural to use the potential outcomes framework (e.g. Imbens and Wooldridge 2009) to phrase our questions about the estimands of interest.

For each individual i, $i = 1, ..., N$ we define two potential outcomes $Y_i(0)$ and $Y_i(1)$.

Outcome $Y_i(0)$ is the outcome individual i would experience if she is not exposed to treatment, similarly $Y_i(1)$ is the outcome she would experience if exposed to treatment. We only observe individual i in or out of treatment so one of the outcomes is a counterfactual outcome. Let W_i be an indicator that takes on the value 1 if i is treated and 0 otherwise. Then we can write the observed outcome Y_i:

$$Y_i = Y_i(0)(1 - W_i) + Y_i(1)W_i$$

Our interest centers around the population average treatment effect (or some conditional version there of): $E[Y_i(1) - Y_i(0)]$. The "problem" with this expression is that we never observe both outcomes for the same individual. To see how we get identification from notice that we can write this expression:

$$\begin{aligned} E[Y_i(1) - Y_i(0)] &= P(W_i = 0)E[Y_i(1) - Y_i(0)|W_i = 0] \\ &\quad +(1 - P(W_i = 0))E[Y_i(1) - Y_i(0)|W_i = 1] \\ &= P(W_i = 0)(E[Y_i(1)|W_i = 0] - E[Y_i(0)|W_i = 0]) \\ &\quad +(1 - P(W_i = 0))(E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 1]) \end{aligned}$$

the first and last conditional expectation in the last expression above is unobserved. We get identification from the experimental design in particular since assignment is independent of potential outcomes: $W_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ for all i, which implies that $E[Y_i(j)|D_i = 0] = E[Y_i(j)|D_i = 1]$, $j = 0, 1$, so that we may write:

$$E[Y_i(1) - Y_i(0)] = E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 0] \tag{1}$$

Equation (1) answers the question: what is the effect of treatment on the population of applicants? Additionally since we are interested in the question how treatment affects the two groups we can decompose the equation further. Let G_i denote the group membership of individual i. We set G_i equal to 1 if i is a majority applicant, and 0 otherwise.

$$\begin{aligned} E[Y_i(1) - Y_i(0)] &= Pr(G_i = 1)E[Y_i(1) - Y_i(0)|G_i = 1] \\ &\quad +(1 - Pr(G_i = 1))E[Y_i(1) - Y_i(0)|G_i = 0] \end{aligned}$$

where again identification follows from random assignment.

While the estimands can be estimated directly from sample averages, it will be useful (as will become clear later) to explicitly model potential outcomes as follows. Suppose that the potential outcome in the absence of treatment can be written as:

$$Y_i(0) = \alpha + \delta G_i + \epsilon_i$$

where $\alpha \geq 0$ is a constant, G_i denotes group membership, and $G_i = E[Y_i(0)] - Y_i(0)$, i.e. the deviation of individual i's outcome from the population average. Suppose in addition that treatment effect is:

$$Y_i(1) - Y_i(0) = \tau + \tau_1 G_i$$

This specification assumes constant treatment effect across individuals, but allows for a group specific treatment effect τ_1. The observed outcome is $Y_i = Y_i(0)(1 - W_i) + Y_i(1)W_i$. Substituting yields the regression equation:

$$Y_i = \alpha + \delta G_i + \tau W_i + \tau_1 G_i W_i + \epsilon_i$$

4.4 Estimates of Unconditional Average Treatment Effects

Here we present raw estimates of the treatment effects without correcting for covariates, and other features of the data.

Table 6 looks at the interview decision. The table is broken down by group membership and further broken down by the educational requirements of the vacancy. For the full sample of majority applicants 15.7% of those in the control group reached the interview stage as opposed to 9.6% in the treatment group. This difference is significant at the 1% level (p=.0038). Decomposing further on educational requirement we see that the treatment effect is negative for the low and mid level vacancies (673 and 207 observations respectively) whereas its positive for high level jobs (114 observations). For minority candidates the estimated average treatment effect is small and insignificant. 8.7% of minority candidates in the control group reached the interview stage, whereas 8.8% in the treatment group reached the interview stage.

In table 7 the outcome variable is whether the applicant receives a job offer or not. The sample consists of all applicants. 7.4% of majority applicants in the control group receive a job offer, while 3% in the treatment group receives an offer. The treatment effect

Group	Education Entry Requirement	Interview		
		Control	**Treatment**	**Difference**
Majority	Low	.1408	.0833	-.0575**
	Middle	.2462	.0915	-.1546***
	High	.1429	.3478	.2050**
	Total	.1570	.0963	-.0608***
Minority	Low	.0633	.0723	.0090
	Middle	.2	.1538	-.0462
	High	.1	.1667	.0667
	Total	.0865	.0882	.0017

Notes: p-value for test of equal proportions. *** p <0.01, ** p <0.05, * p <0.1

Table 6: Estimated Average Treatment Effects on Interview Invitation. Conditional on Group and Job Requirement.

is significant at the 1% level (p= .0016). Turning to the minority candidates there is no evidence of an effect of treatment.

Group	Education Entry Requirement	Offer		
		Control	**Treatment**	**Difference**
Majority	Low	.0794	.0253	.0542***
	Middle	.0923	.0352	.0571*
	High	.0440	.0870	-.0430
	Total	.0739	.0303	.0436***
Minority	Low	.0380	.0241	.0139
	Middle	.0667	0	.0667
	High	0	.1667	-.1667
	Total	.0385	.0294	.0090

Notes: p-value for test of equal proportions. *** p <0.01, ** p <0.05, * p <0.1

Table 7: Estimated Average Treatment Effects on Offer Decision. Conditional on Group and Job Requirement.

Table 8 presents estimates of some key parameters of interest. For each outcome of interest (interview invitation and job offer) the table presents estimates for δ, the extent of differential treatment in the absence of intervention, and τ_1, the group specific effect of treatment[27]. The estimate of δ is obtained as the difference in the sample averages

[27]Estimates are obtained from OLS on the full sample using robust standard errors to correct for the

between the two groups conditional on being in the control. In addition τ_1 is estimated, as a difference-in-difference between treatment effects conditional on group membership.

Based on these raw estimates (which does not control for covariates) the estimate of differential treatment in favor of the majority group is 7.1% (p=.031). There is no statistically significant evidence of a differential treatment effect (p=.164), but the point estimate is -6.3%. Turning to the offer decision the differential treatment at the interview stage maps into a higher offer rate for the majority group, the point estimate is 3.5%, but is not statistically significant (p=.119).

Decision	Parameter	Estimate
Interview	δ	.0705**
	τ_1	-.0625
Offer	δ	.0354
	τ_1	-.0346

Notes: *** p <0.01, ** p <0.05, * p <0.1

Table 8: Estimate of group difference in control condition (δ) and group specific treatment effect (τ_1).

4.5 Regression Estimates

The raw estimates presented above ignore several features of the data. First, applicants decide which vacancy to apply for. Although applicants cannot condition their application decision on whether the vacancy is in the control or treatment group, as this is hidden to them, characteristics may differ systematically between groups. A particular concern is that the difference in success rates in the control condition may be due to majority candidates being better qualified. Note however that this is unlikely to be the case as success rates in landing an interviews is roughly equalized in the treatment condition. Specifically if the majority group is better qualified we would expect this to also be present in the treatment condition.

Second, in the framework introduced above we assumed that the potential outcome in the control group is constant across vacancies: $Y_i(0) = \alpha + \delta G_i + \epsilon_i$. This is unlikely to

heteroskedasticity due to the outcome variable being binary.

be satisfied in our sample as different vacancies may have different likelihoods of success. Specifically vacancies may have different optimal thresholds as in the model of section A. A more practical issue is that some vacancies contain several openings. We allow for a vacancy specific fixed effect, α_j, to address these issues.

Finally, the data is clustered in a specific way. In particular within vacancies there is competition among the applicants for a limited number of interview spots. We would therefore expect that the success of one candidate negatively affects the chances of another candidate to succeed in landing an interview. More generally we allow for correlation within vacancies and cluster errors on the vacancy level.

Using the notation introduced above we model the outcome (interview or offer) of individual i in vacancy j when i is not treated as:

$$Y_{ij}(0) = \alpha_j + \beta X_{ij} + \delta G_i + \epsilon_{ij}$$

where α_j is a vacancy specific effect, β is a vector of parameters and X_{ij} a vector of candidate (age, educational attainment, gender, etc.) and vacancy specific variables (and interactions), and ϵ_{ij} is the mean deviation of the outcome of individual i in job j.

With this model for $Y_{ij}(0)$ we arrive at the regression specification:

$$Y_{ij} = \alpha_j + \beta X_{ij} + \delta G_i + \tau_1 G_i \times W_j + \epsilon_{ij} \tag{2}$$

where as before G_i is a group membership dummy which takes the value 1 if i is a majority member, and 0 otherwise. W_j is a treatment dummy which takes the value 1 if vacancy j is assigned to treatment and 0 otherwise.

Since we allow for a vacancy specific effect α_j we cannot separately identify the overall effect of treatment (τ).

4.5.1 Estimation: The Interview Decision

The specification in equation (2) readily suggests an estimation procedure. Since the outcome variable is dichotomous this specification is a linear probability model (LPM) estimated by OLS[28].

[28] As noted in section 4.2.2 there are 51 candidates with missing information for either their educational level or their age. We drop these observations from the estimation sample. Accordingly the estimation sample contains 1149 individuals, applying for a total of 37 vacancies.

Table 9 presents a series of estimates based on the specification given in equation (2)[29]. Model (1) is a basic specification which only includes the group membership dummy, the treatment dummy and their interaction, and vacancy specific fixed effects. Holding membership in the majority group increases the likelihood of success by 7.7%. This effect is significant at the 5% level. Model (2) allows for candidate and vacancy covariates (and interactions), but does not allow for vacancy specific effects. Not all parameter estimates are reported, but those for gender and age are presented since some empirical literature has identified differential treatment based on gender and age. We do not find evidence of such discrimination. Introducing covariates improves the precision of the estimate for differential treatment (significant at 1% level). The size of the effect is similar (8.2%). In this specification there is also (weak) evidence that the two groups are affected differentially by treatment (10% level). Finally model (3) allows for vacancy specific effects and covariates. This specification subsumes some of the vacancy specific covariates which cannot be separately identified as they are constant within a vacancy.

In the last row of the table we report the p-value for the test $H_0: \ \delta + \tau_1 = 0$. In all specifications the policy is successful in the sense that we cannot reject that treatment succeeds in creating equal chances.

A potential concern when interpreting results is omitted variables. Our data only contain a limited set of covariates. Recruiters will have more detailed available to them when making a decision whether to interview or not. The application package for a candidate includes a cover letter and a detailed resume, whereas we only have access to a few candidate characteristics. It is plausible that these characteristics (unobserved to us) are systematically different between groups. E.g. the population of majority members have a stronger labor market attachment, and thus one would expect that the population of majority members have more labor market experience and perhaps better references. Such characteristics are a strong signal of productivity. If this population statement is also true for the population of applicants then, since we cannot condition on these variables, this

[29] Appendix C compares this specification with estimates obtained from a probit and fixed effects probit specification. Estimated marginal effects and significance levels remain stable relative to the LPM model. In addition we also tried a specification that allows for heterogeneous effects based on the educational requirement of the job, which we return to in section 5. We found no evidence of differences in differential treatment based on job categories (available upon request).

correlation will be picked up by the group membership variable. This concern is unlikely to be the driver of our results. Recruiters in the treatment condition also have access to detailed information on labor market experience, but group membership is hidden. Thus we would expect to find that group membership matters for the selection decision also in the treatment condition. As it turns out there is no evidence of differential treatment in the treatment condition. Thus at least for the present sample omitted variables correlated with group membership cannot be the driver of the results.

		Dependent Variable: Interview		
		(1)	(2)	(3)
Majority (δ)		0.0765**	0.0818***	0.0873***
		(0.0317)	(0.0263)	(0.0300)
Majority×Treatment (τ_1)		-0.0575	-0.0649*	-0.0556
		(0.0408)	(0.0365)	(0.0398)
Female			0.0155	0.0398
			(0.0219)	(0.0250)
Age	30-45		0.0243	0.0328
			(0.0257)	(0.0244)
	45-65		-0.0451	-0.0314
			(0.0312)	(0.0319)
Observations		1,149	1,149	1,149
Clusters		37	37	37
Job FE		YES	NO	YES
Matchspecific Controls		NO	YES	YES
R-squared		0.119	0.058	0.151
$\delta + \tau_1 = 0$ (p-value)		0.462	0.522	0.273

Notes: Clustered standard errors (per vacancy) in parentheses.
*** p <0.01, ** p <0.05, * p <0.1. Excluded age category is < 30 years of age. Model (2): 12 observations with predicted probability less than 0. No observations with predicted probability greater than 1. Model (3): 5 observations have a predicted probability less than 0. No observations with a predicted probability greater than 1.

Table 9: LPM estimates for Interview Decision

To put the estimated coefficients (from model (3)) in perspective we fix characteristics at their sample means. A minority candidate with average characteristics in the control group has a likelihood of success of 7.3%. Their majority counterpart has a chance of success of 16%. Thus while a minority candidate would expect to send around 13-14 applications before landing an interview the majority candidate can expect to send only 6 applications.

4.5.2 Estimation: The Offer Decision

The estimates for the decision to interview confirm that group membership plays a significant role for the success of candidates in landing an interview. We now look at the job offer rates.

The regressions reported in table 10 are identical to those reported for the interview decision outcome above, except that the outcome (dependent) variable is a binary indicator of whether the candidate was offered a job or not.

Model (1) includes only the group dummy, it's interaction with the treatment and the vacancy specific effects. The point estimate of differential treatment is 2.9% but the effect is not statistically significant at conventional levels. In model (2) we leave out the vacancy specific effects but introduce candidate and job-covariates. The point estimate of differential treatment in this specification is 4.2% and it is statistically significant at the 5% level. Model (3) adds vacancy specific effects. The point estimate now drops to 3.7% and is no longer significant at conventional levels. In terms of fit model (2) and (3) for the offer decision has a relative large number of candidates for which the model predicts probabilities below 0[30].

In sum we find only weak evidence of differential treatment in unconditional offer rates. Taken together with the evidence of differential treatment in the interview decision these results are compatible with minority candidates being held to a higher standard in the interview decision, but conditional on making it to the interview stage minority candidates perform better than their majority counterparts. In the light of differential treatment at the interview stage one would indeed expect the pool of minority candidates to be on average better qualified than the interview pool of majority candidates, which should translate into higher offer rates. Due to the relatively low sample size of candidates making it to the interview stage the data does not allow us to precisely estimate this effect (regressions for

[30] In appendix C we fit a probit model to the data. In some of these specifications the majority indicator is significant at the 10% level.

the pool of applicants who make it to the interview stage can be found in Appendix B).

		Dependent Variable: Job Offer		
		(1)	(2)	(3)
Majority (δ)		0.0286	0.0415**	0.0365
		(0.0249)	(0.0200)	(0.0239)
Majority×Treatment (τ_1)		-0.0280	-0.0339	-0.0279
		(0.0294)	(0.0248)	(0.0285)
Female			0.00271	0.0123
			(0.0154)	(0.0176)
Age	30-45		0.00586	0.00778
			(0.0152)	(0.0147)
	45-65		-0.0253	-0.0260
			(0.0166)	(0.0181)
Observations		1,149	1,149	1,149
Clusters		37	37	37
Job FE		YES	NO	YES
Matchspecific Controls		NO	YES	YES
R-squared		0.083	0.042	0.102
$\delta + \tau_1 = 0$ (p-value)		0.969	0.616	0.599

Notes: Clustered standard errors (per vacancy) in parentheses. *** p <0.01, ** p <0.05, * p <0.1. Excluded age category is < 30 years of age. Model (2): 77 observations with predicted probability less than 0. No observations with predicted probability greater than 1. Model (3): 83 observations have a predicted probability less than 0. No observations with a predicted probability greater than 1.

Table 10: LPM Estimates Offer Decision. Full Sample.

4.6 Which Discrimination?

Does the experiment allow us to shed light on which theory of differential treatment better is in line with our results?

Two prominent (general equilibrium) economic theories of group based differential treatment exist. The animus-based theory due to Becker (1957) assumes that a recruiter has a taste for discrimination. The taste may come from a pure dis-taste of hiring minority members or a taste for hiring majority members (or more generally the recruiter may internalize dis-taste from her current employees). In the other prominent theory the source of differential treatment is informationally based (Arrow 1973, Phelps 1972). Because recruiters cannot observe all relevant productivity characteristics at the application stage they may form stereotypes based on group membership.

The present experiment is not specifically designed to distinguish between these two theories. In a recent paper List (2004) used a field experiment coupled with lab experiments on the same population. His setting, unlike ours, is trade at sport cards fares where in addition to taste based and statistical discrimination, also differences in bargaining techniques may explain different group outcomes. List finds overwhelming support for the predictions of statistical discrimination. He is able to uncover that different consumer groups have different value distributions, and that experienced sellers are aware of this and use this knowledge to third degree price discriminate buyers. We are not aware of studies similar studies in a labor market setting. Returning to the present setting the fact that the recruitment procedure is a two-step procedure, and that we observe how recruiters behave in frames where they can observe group memberships and when they cannot (until the interview stage) allow us to make some indirect inference about which of these theories is more consistent with observed recruitment decisions.

The control group behavior cannot be used to distinguish between the theories, as both would predict that minority candidates are held to a higher standard. In the animus based theory better productivity characteristics of a minority candidate compensates the employer for her dis-taste of minority members. In the informationally based theory higher standards for the minority group is set since employers believe that the group on average is less qualified than the majority group. A minority applicant therefore needs to be more convincing to land an interview.

Behavior in the treatment condition is more in line with an informationally based theory of differential treatment. In particular equally qualified candidates face the same chances of landing an interview in the treatment condition, but also at the second stage there is no evidence of differential treatment. Table **??** in the appendix contains regression results for the offer decision conditional on reaching the interview stage. The point estimate for the effect of group membership actually goes in the direction of an advantage to being a minority member, but the effect is not statistically significant. Under an animus-based theory we would expect discrimination at the offer stage in the treatment group. If recruiters cannot express their taste at the first stage of the application procedure they are able to do so at the interview stage. On the other hand a theory based on statistical discrimination would predict that the information contained in group membership is used at the first stage to set tougher standards for minorities. Therefore at the interview stage where all information is revealed we would not expect group membership to matter for predicting receipt of an offer.

Some of our results are somewhat puzzling. E.g. we observe that majority membership confers an advantage for the interview decision, but no such advantage is identified for unconditional job offers. If recruiters used information optimally (in particular information on group membership), at the interview decision stage, we would expect that conditional on getting an interview no differences be detected between groups. This suggests that minority candidates are penalized "too much" based on group membership in the interview decision. Thus the differential treatment that we observe seems partially to reflect recruiters knowledge of the distribution of unobserved (at the application stage) characteristics and the optimal response to this.

5 A Follow Up Experiment

The experimental results appear to us to give strong and conclusive evidence about differential treatment in the decision to interview based on the classification of applicants into minority and majority membership. Interestingly this differential treatment does not translate into differential treatment in job offers.

The policy maker's stated pre-experimental objective was to: "examine if minorities have a higher chance to be invited for interview under an anonymous application procedure"(Gemeente Nijmegen 2008, p.3). Thus the policy maker was first and foremost concerned whether a "bias" was present in the first stage of the application procedure. Whether differential treatment was present for the decision to offer a job did not enter explicitly into the policy maker's objective. The method of evaluation of the experimental data differed from the one we have employed here. To be specific the success of the policy was evaluated by comparing the percentage of minority (majority) applicants invited for an interview under treatment and control.

Under the policy makers evaluation method significant levels of differential treatment were identified. In spite of this concerns were raised about reliability. The following quote (translated from Dutch) illustrate some of these (ex-post) concerns (published in March 2007):

> When evaluating the first experiment we find that applicant characteristics such as age, gender and educational level does not, in a statistical sense, contribute to explaining which candidates are invited. It could be that subjective factors played a role when recruiters select candidates. We are thinking of differences in the way letters are assessed such as preferences for a certain letter style, references provided in the letter, and preferences for candidates based on (un)-conscious stereotyping . Moreover the results could be due to chance, maybe the difference would also have occurred in the absence of an experiment. Finally, it may be that recruiters paid more attention to the ethnicity of the candidate because of the experiment. (Gemeente Nijmegen 2008, p.4-5)

In April 2007 it was decided that a follow up experiment to test the reliability of the first experiment be conducted. The follow up ran from 1st of May 2007 until 1st of January 2008, and involved the same departments as in the first experiment. The twist was that the former control group would become the treatment group and vice versa (it was still not made public which departments had been assigned to treatment control in the first experiment).

This follow up experiment provides us with an opportunity to understand how recruiters respond to the announcement. Before proceeding it is worthwhile to raise a number of ex-ante concerns about the interpretation of a follow up experiment on the same population. A main concern is that those that were treated in the first experiment become the control in the second experiment. If recruiters, and particularly recruiters in the former treatment group, take the participation in an experiment as a signal of the objective function of the council then at least one would expect that a sophisticated manager could surmise that his/her decisions would be monitored. Decisions that are not aligned with the principal may therefore lead to sanctions.

An interesting feature of the follow up is that the informational setting has changed. In particular the method of evaluation has become public knowledge; prior to the first experiment it is not clear that recruiters were aware how their decisions would be scrutinized, and which method would be used to identify differential treatment. After the publication of the evaluation of the first experiment it is now clear that the experimental outcome is evaluated by comparing (unconditional) invitation rates for the two groups. Differential treatment is present if invitation rates are significantly different in the control condition. Recall that in the first experiment the invitation rate for the majority group was 16% and 10% for the minority group in the control treatment. This difference is statistically significant (p=.032). On the other hand in the treatment condition invitation rates are 10% and 9% respectively (p=.794).

We do not explicitly know the aim of recruiters, but we do know from survey evidence that many recruiters expressed reservations about the policy measure. In order to avoid the policy becoming permanent invitation rates in the control treatment needs to be equalized. How should a recruiter who harbors stereotypes respond to the information that has been revealed? This question can be analyzed within the context of the interview decision model we introduced in appendix A. Suppose for concreteness that jobs are created (exogenously) at two job levels: H(igh) and L(ow); the H job level requires a higher standard being set in the unconstrained problem. In appendix A.3 we show that in general it is not optimal to respond to the equal invitation rate requirement by equalizing invitation rates between groups at each job level. The optimal way to respond depends on two factors. First, when the share of minority applicants is small, the recruiter responds to the constraint by distorting standards more for the minority group than for the majority group (i.e. lowering standards for the minority group and increasing standards for the majority group). This effect is the standard effect identified by Coate and Loury (1993). The intuition is that with a small minority share distorting the minority threshold away from the optimum, in order to meet the constraint, is less costly than distorting the majority threshold, since the recruiter only bears this cost for a small fraction of the total applicant pool. Thus the recruiter distorts the minority threshold relative more than the minority threshold. A second effect that we identify concerns the question regarding at which job level the recruiter distorts standards relatively more. This turns out to depend on the density of applicants around the interview cutoffs. For a fixed distortion in the cutoff the higher the density of applicants around the cutoff the larger the effect in terms of meeting the constraint. For example if the density around the L job cutoff is larger than at the H job, then ceteris paribus the optimal distortion at the L job will be greater[31] These two consideration may lead to heterogeneous effects. In particular whereas the unregulated solution displays higher cutoffs for minorities at both job levels relative to the tier specific majority cutoffs, the regulated solution may display a lower cutoff for minorities at the L level but a higher cutoff at the H level.

To investigate this hypothesis we modify the framework to allow for heterogeneous effects. The sample contains 3 categories of jobs based on the educational requirement ("low", "middle" and "high"). Label these r=1, 2 and 3 respectively, 3 corresponding to the "high" requirement. Let F_j^r be an indicator that takes value 1 if vacancy j has educational requirement equal to r and 0 otherwise. We can then allow for heterogeneous differential treatment in the control condition as follows:

$$Y_{ij}(0) = \alpha_j + \sum_{r=1}^{3} \delta_r G_i F_j^r + \epsilon_{ij}$$

The treatment effect is modeled as: $Y_{ij}(1) - Y_{ij}(0) = \tau_0 + \sum_{r=1}^{3} \tau_r F_j^r G_i$. Then the regression equation becomes (ignoring covariates):

$$Y_{ij} = \alpha_j + \delta_1 G_i + \sum_{r=2}^{3} \delta_r F_j G_i + \tau_0 W_j + \tau_1 G_i W_j + \sum_{r=2}^{3} \tau_r G_i F_j^r W_j + \epsilon_{ij} \tag{3}$$

In this specification δ_1 is a measure of differential treatment in "low" educational requirement jobs, $\delta_1 + \delta_2$ is the corresponding measure for "middle" range jobs, and $\delta_1 + \delta_3$ measure differential treatment in the "high" range jobs. As before when estimating the model the vacancy fixed effects subsumes τ_0.

5.1 Results from the Follow Up

As noted above the follow up experiment ran from 1st of May 2007 until 1st of January 2008. The experimental sample contains a total of 48 openings for which 1393 applicants applied. In the interest of brevity the descriptive statistics of the sample have been relegated to

[31]A sufficient condition for this is $f_q'(\theta) < 0$. The example in the appendix shows that this condition is not necessary.

appendix D. Instead we proceed directly to estimating specifications along the lines of equations (2) and (3).

Table 11 contains estimation results for 4 different specifications for the interview decision[32]. Specifications (1) and (2) in the table assumes treatment is constant across job categories. Model (2) differs from (1) in that we allow for candidate and vacancy specific covariates. All specifications include vacancy specific fixed effects. Specification (1) is a simple specification which includes the group indicator and treatment indicator and their interaction. There is no evidence of differential treatment in this specification. Specification (2) also includes candidate covariates and interaction terms between the educational level of the candidate and the required educational level (labeled matchspecific controls). Also in this specification there is no evidence of differential treatment. There is some evidence of candidates in the middle age category having an advantage viz-a-viz candidates below the age of 30. Specification (3) leaves out candidate and matchspecific covariates, but allows for heterogeneous effects. For job categories in the mid range the estimate of differential treatment based on group membership is .1052 (p = .0948). For top range jobs the point estimate increases to .2484 (p = .0171). Specification (4) allows for both heterogeneous effects and applicant covariates. Also in this specification significant differences in differential treatment per job category are present. The point estimate in the middle job category is .1371 (p = .0949). For the high range jobs the estimate increases to .2435 (p = .0128).

Table 12 presents estimation results for (unconditional) offer decision based on the same specifications as in table 11. None of the specifications indicate significant differences based on group membership. Therefore it appears, as it was the case in the main experiment, that the differential treatment in the interview decision does not carry over to the offer decision.

6 Discussion

We have presented evidence from a field experiment on the role of group membership in interview and job offer decisions of recruiters in the public sector. To our knowledge this is

[32]The estimation sample contains 1324 observations due to missing values on applicant covariates. Estimating model (1) and (3), which do not contain covariates, on the full sample yields identical results.

		Interview Decision			
		(1)	(2)	(3)	(4)
Majority	(δ_1)	-0.0180	0.00549	-0.0718	-0.0442
		(0.0417)	(0.0366)	(0.0554)	(0.0471)
Majority×	(δ_2)			0.177**	0.181*
Middle Education Job				(0.0829)	(0.0925)
Majority×	(δ_3)			0.364***	0.288***
High Education Job				(0.131)	(0.107)
Majority×	(τ_1)	-0.0239	-0.0203	0.0258	0.0303
Treatment		(0.0592)	(0.0526)	(0.0966)	(0.0874)
Majority×Treatment×	(τ_2)			-0.166	-0.176
Middle Education Job				(0.122)	(0.123)
Majority×Treatment×	(τ_3)			-0.382	-0.344
High Education Job				(0.247)	(0.234)
Female			0.0209		0.0216
			(0.0301)		(0.0301)
Age	30 – 45		0.0753**		0.0784**
			(0.0343)		(0.0342)
	45 – 65		-0.0498		-0.0457
			(0.0362)		(0.0363)
Observations		1,324	1,324	1,324	1,324
Clusters		48	48	48	48
R-squared		0.112	0.160	0.116	0.162
Job FE		YES	YES	YES	YES
Matchspecific Controls		NO	YES	NO	YES
$\hat{\delta}_1 + \hat{\delta}_2$				.1052*	.1371*
$\hat{\delta}_1 + \hat{\delta}_3$				.2484**	.2435**
H_0: Equal Treatment under AAP (p-value)				0.771	0.975

Notes: Clustered standard errors (per vacancy) in parentheses. *** p <0.01, ** p <0.05, * p <0.1. Excluded age category is < 30 years of age.

Table 11: LPM Estimates for the Interview Decision for Follow Up Experiment. Full sample.

		Offer Decision			
		(1)	(2)	(3)	(4)
Majority	(δ_1)	-0.0161	-0.00833	-0.0254	-0.0170
		(0.0246)	(0.0221)	(0.0298)	(0.0263)
Majority×	(δ_2)			0.0152	0.0202
Middle Education Job				(0.0675)	(0.0684)
Majority×	(δ_3)			0.0930*	0.0731
High Education Job				(0.0554)	(0.0485)
Majority×	(τ_1)	-0.00389	-0.00165	0.00904	0.0114
Treatment		(0.0332)	(0.0314)	(0.0462)	(0.0429)
Majority×Treatment×	(τ_2)			-0.0010	-0.0152
Middle Education Job				(0.0810)	(0.0799)
Majority×Treatment×	(τ_3)			-0.178	-0.167
High Education Job				(0.128)	(0.115)
Female			0.0210*		0.0210
			(0.0125)		(0.0126)
Age	30 – 45		0.0208		0.0212
			(0.0156)		(0.0156)
	45 – 65		-0.0139		-0.0134
			(0.0164)		(0.0168)
Observations		1,324	1,324	1,324	1,324
Clusters		48	48	48	48
R-squared		0.051	0.066	0.052	0.068
Job FE		YES	YES	YES	YES
Matchspecific Controls		NO	YES	NO	YES
H_0: Equal Treatment under AAP (p-value)				0.745	0.806

Notes: Clustered standard errors (per vacancy) in parentheses. *** p <0.01, ** p <0.05, * p <0.1. Excluded age category is < 30 years of age.

Table 12: LPM Estimates for the Offer Decision for Follow Up Experiment. Full sample.

the first study that credibly estimates the effect of group membership in both the interview and offer decision.

We find strong evidence that majority membership confers an advantage in reaching the interview stage. Moreover the effect is sizeable. Correspondence studies have become popular since they rely on fictitious resumes, and therefore leaves all aspects of the information transmitted to recruiters under the control of the experimenter. Market sorting aside this methodology yields credible estimates of differential treatment. However it can be argued that the outcome variable of interest is job offers not the interview decision. An advantage of the present study is that we can also observe this second stage. Perhaps surprisingly we do not find evidence of differential treatment in the job offer decision. It is important to emphasize that we do not get this for free. To obtain these estimates we are giving up substantial control over the information available to recruiters about candidates. But since we do observe how candidates perform also when recruiters are barred from observing group membership, it seems at least in the present sample that no biases have been introduced as a result of this. We find weak evidence that those minority candidates that are selected into the interview pool have better chances than their majority counterparts of receiving a job offer. This suggests that recruiters place too much emphasis on the group variable for the decision to interview. However the low sample size implies that such a conclusion can be at best tentative and certainly deserves more attention.

Turning to the follow up experiment we find evidence that is in line with recruiters reacting strategically to the constraint imposed upon them by the principal. Survey evidence conducted among recruiters identified two main features: (1) recruiters did not believe that they were discriminating, (2) they were skeptical about the use of anonymous application procedures. That recruiters are skeptical about anonymous procedures is perhaps less surprising given that they do not believe that they apply different standards in the first place. We found evidence consistent with the idea that recruiters use preferential treatment in so far as they can achieve equality in invitation rates.

Although we have emphasized identification of differential treatment, the paper also contains results of interest to policy makers. In particular the study contains an evaluation of a particular variant of anonymous application procedures. This procedure seems to work well in creating equal chances in our sample. While we have identified a clear benefit, the policy may also contain substantial costs. An anonymous application procedure hides pay-off relevant information from recruiters. Recruiters may e.g. respond to this constraint by having a larger pool of invited applicants which increases resources dedicated to the hiring procedure. A more noisy first stage may also decrease the quality of the match. Such effects are likely to show up in the longer term in terms of increased separation probabilities. While these are interesting questions in their own right they are beyond the scope of this study.

This experiment is conducted in a particular segment of the labor market, which warrants some comments about external validity. This field experiment is carried out in the public sector. The private market may function differently from the public sector, arguably the private sector may be more disciplined by decisions that do not maximize profits, although as noted by Arrow (1973) this is a delicate argument. Indeed Bertrand and Mullainathan (2004) respond to wanted ads from a large number of private firms in Chicago and Boston, using fictitious resumes. They also identify substantial differential treatment for the decision to interview in this private sector setting.

Another word of caution concerns the nature of the vacancies. Vacancies in the main experiment is heavily concentrated at the lower and middle range. Accordingly the main experiment cannot credibly estimate differential treatment in jobs that require studies at university level. In the follow up the composition of vacancies is more slanted towards higher end jobs but as pointed out the nature of the follow up raises methodological concerns.

Many interesting questions remain. E.g. it appears that recruiters place too much emphasis on group membership when selecting whom to interview, since we find differential treatment at the interview stage but not in the full sample job offer decision. Accordingly a minority candidate selected for interview is on average a "better" match than the majority candidate. This suggest that part of the differential treatment in the interview decision is due to a bias in favor of majority candidates. Arguably the two stages of the application procedure differ substantially in the ease with which a bias can be expressed. A recruiter sorting through resumes may use heuristics when sorting through a large number of applications, but devote more effort and attention at the interview stage where typically an interview committee is present. While our study cannot address the behavioral mechanisms behind the experimental results these are interesting avenues to explore.

References

Altonji, J. G., and R. M. Blank (1999): "Chapter 48 Race and gender in the labor market," vol. 3 of *Handbook of Labor Economics*, pp. 3143–3259. Elsevier.

Arrow, K. J. (1973): "The Theory of Discrimination," in *Discrimination in Labor Markets*, ed. by O. Ashenfelter, and A. Rees, pp. 3–33. Princeton University Press, Princeton, NJ.

Åslund, O., and O. Skans (2007): "Do anonymous job application procedures level the playing field?," IFAU - Institute for Labour Market Policy Evaluation / Working Paper Series.

Becker, G. S. (1957): *The Economics of Discrimination*. University of Chicago Press, Chicago, IL.

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

Bovenkerk, F., M. Gras, D. Ramsoedh, M. Dankoor, and A. Havelaar (1995): "Discrimination against migrant workers and ethnic minorities in access to employment in the Netherlands," *International Migration Papers 4*.

Carlsson, M., and D.-O. Rooth (2007): "Evidence of ethnic discrimination in the Swedish labor market using experimental data," *Labour Economics*, 14(4), 716–29.

——— (2008): "Is It Your Foreign Name or Foreign Qualifications? An Experimental Study of Ethnic Discrimination in Hiring," IZA Discussion Paper No. 3810.

Coate, S., and G. C. Loury (1993): "Will Affirmative-Action Policies Eliminate Negative Stereotypes?," *American Economic Review*, 83(5), 1220–40.

Eriksson, S., and J. Lagerström (2007): "Detecting Discrimination in the Hiring Process: Evidence from an Internet-based Search Channel," IFAU Working Paper No. 2007:19, Institute for Labour Market Policy Evaluation, Sweden.

Gemeente Nijmegen (2008): "Rapportage 'anonimiseren van sollicitatiebrieven': Evaluatie eerste en tweede meting," Nijmegen: Gemeente Nijmegen.

Goldin, C., and C. Rouse (2000): "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, 90(4), 715–41.

Harrison, G., and J. List (2004): "Field experiments," *Journal of Economic Literature*, 42(4), 1009–1055.

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

Imbens, G., and J. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Kruisbergen, E., and T. Veld (2002): "Een gekleurd beeld. Over beoordeling en selectie van jonge allochtone werknemers," Assen: Koninklijke Van Gorcum.

List, J. A. (2004): "The Nature and Extent of Discrimination in the Marketplace: Evidence From the Field," *Quarterly Journal of Economics*, 119(1), 49–89.

Lundberg, S. J., and R. Startz (1983): "Private Discrimination and Social Intervention in Competitive Labor Markets," *American Economic Review*, 73(3), 340–47.

Neal, D. A., and W. R. Johnson (1996): "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, 104(5), 869–95.

Norman, P. (2003): "Statistical Discrimination and Efficiency," *Review of Economic Studies*, 70(3), 615–627.

Phelps, E. S. (1972): "The Statistical Theory of Racism and Sexism," *American Economic Review*, 62(4), 659–61.

Riach, P. A., and J. Rich (2002): "Field Experiments of Discrimination in the Market Place," *Economic Journal*, 112(483), F480–F518.

Scheepers, P., R. Eisinga, and L. Linssen (1994): "Etnocentrisme in Nederland: Veranderingen bij kansarme categorieën?," *Sociologische Gids*, 3, 185–201.

Verberk, G., P. Scheepers, and A. Felling (2002): "Attitudes and behavioural intentions towards ethnic minorities. An empirical test of several theoretical explanations for the Dutch case," *Journal of Ethnic and Migration Studies*, 2, 197–217.

A A Simple Model for the Decision to Interview

In this section we present a simple model for the decision to interview. The model follows Coate and Loury (1993).We show how the model specification lends itself naturally to a regression interpretation.

A risk neutral employer decides whether to invite candidate i to interview for job j. The applicant is either a good match (q) or a bad match (u) for the job. The employer cannot directly observe the quality of the match, but she receives an informative signal, θ, about the suitability of the candidate for job j. If the match is good then the signal is drawn from a distribution with density $f_q(\theta)$, and otherwise the signal is drawn from a distribution with density $f_u(\theta)$. Both conditional densities are continuous and have full support on the real line. We assume that the ratio $f_q(\theta)/f_u(\theta) \equiv g(\theta)$ satisfies the monotone likelihood ratio property[33]. In other words the higher the signal the more likely the candidate is to be a good match for the job. The employer has belief $0 < \pi^0 < 1$ that the candidate is a good match. After receiving the signal the employer updates her prior belief using Bayes' law:

$$\pi^*(\theta) = \frac{\pi^0 f_q(\theta)}{\pi^0 f_q(\theta) + (1 - \pi^0) f_u(\theta)}$$

To solve for the optimal decision rule we need to specify the payoffs for the employer. Assume that a candidate who is invited and is a good match yields a payoff of $\chi_q > 0$ and $-\chi_u < 0$ otherwise. If the employer does not invite the candidate her payoff is normalized

[33]This ensures that the employers optimal decision rule is a cut-off rule.

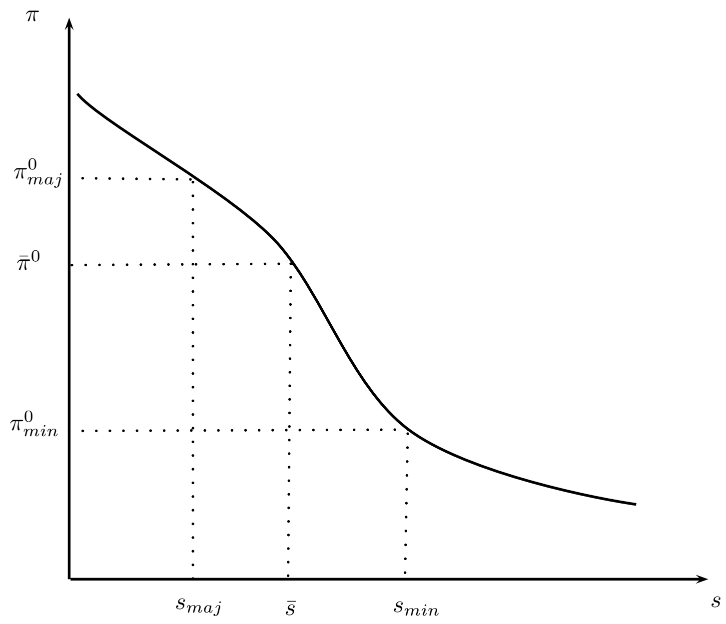


Figure 5: Optimal Standards

to 0. Thus the candidate is invited provided that:

$$\pi^*(\theta)\chi_q - (1 - \pi^*(\theta))\chi_u \geq 0 \tag{4}$$

To see that the optimal rule is a cut-off rule observe that:

$$\frac{\pi^*(\theta)}{1 - \pi^*(\theta)} = \frac{f_q(\theta)}{f_u(\theta)} \frac{\pi^0}{1 - \pi^0}$$

which is monotonically increasing in θ. Now note that equation 4 can be rewritten:

$$\frac{\pi^*(\theta)}{1 - \pi^*(\theta)} \geq \frac{\chi_u}{\chi_q}$$

Thus there exists a unique θ^*, which we call a *standard* such that the candidate is invited if and only if she meets this standard, $\theta \geq \theta^*$.

The workings of the model is illustrated in figure 5. The employers prior belief, π is on the vertical axis and the standard, s, is on the horizontal axis. For a given prior belief the curve identifies the standard (θ^*) such that when the candidate "emits" this signal the employer is exactly indifferent between interviewing and not interviewing the candidate.

Here there are two different groups a majority and a minority group. The recruiter has a pair of prior beliefs (π^0_{maj}, π^0_{min}) about the match quality of candidates. In the figure

$0 < \pi^0_{min} < \pi^0_{maj} < 1$ such that the employer has positive stereotypes about the majority group relative to the minority group. The model then implies that minority candidates face tougher standards $s_{min} > s_{maj}$.

A.1 Anonymous Application Procedures

The model lends itself easily to the analysis of a policy intervention such as the anonymous application procedures. An effective anonymous application procedure (AAP) forces the recruiter to having one standard for all groups. How should the employer respond when she recruits under AAP? Since she can no longer observe group membership she should assign probabilities based on population frequencies (assuming that the policy itself does not change the composition of applicants). If she has a pair of prior beliefs $(\pi^0_{maj}, \pi^0_{min})$, where $0 < \pi^0_{min} < \pi^0_{maj} < 1$ and the population shares are $(p_{maj}, 1-p_{maj})$ then her prior belief that a randomly drawn candidate is qualified, when she cannot observe group membership is just the frequency weighted average:

$$\bar{\pi}^0 = \pi^0_{maj} p_{maj} + \pi^0_{min}(1 - p_{maj})$$

Therefore the population weighted belief is a convex combination of the recruiter's prior beliefs about the two groups. As illustrated in figure 5 the optimal standard, when the recruiter is unable to condition on group membership lies in between the two standards that the recruiter sets when she is able to condition on group membership: $s_{maj} < \bar{s} < s_{min}$. In words under an effective AAP majority candidates will face tougher standards whereas the standard applied to minority candidates will be looser.

A.2 Regression Interpretation of the Basic Model

The model for the interview decision can be given a regression interpretation.

In particular after linearizing, and introducing candidate covariates, X_i, we can think of candidate i generating a score when applying for job j:

$$y^*_{ij} = \tilde{\pi}^0 + \epsilon_{ij}$$

We do not observe the score but we do get to see the decision:

$$y_{ij} = \begin{cases} 1 & if \quad y^*_{ij} \geq 0 \\ 0 & otherwise \end{cases}$$

When estimating the model we allow for a matching function of the form:

$$m(X_i, r_j) \quad = v(X_i, X_j) + \gamma_j$$

where $v(X_i, X_j)$ is a matching function (potentially fully saturated in candidate and vacancy covariates, X_i and X_j respectively) and γ_j is interpreted as a fixed vacancy specific threshold. We write:

$$y^*_{ij} \quad = \tilde{\pi}^0 + v(X_i, X_j) + \gamma_j + \epsilon_{ij}$$

Finally, it is easy to incorporate different group standards into the specification. This is modeled by allowing the employer to have different prior beliefs about the likelihood that a member from a particular group is a good match for the job.

$$y^*_{ij} \quad = \tilde{\pi}^0 + \tilde{\pi}^0_{Min} D_i + v(X_i, X_j) + \gamma_j + \epsilon_{ij}$$

where $\tilde{\pi}^0_{Min}$ is a coefficient that measures the difference between the prior beliefs of the recruiter for the two groups, and D_i is a minority membership dummy.

A.3 The Recruiter's Problem under Regulation

This section considers how a recruiter should respond to a regulator requiring that invitation rates are equalized between groups.

Suppose there are two groups A and B. The total measure of population is normalized to 1, and group A is share $0 < \lambda < 1$. Given prior belief π the expected profits of using standard s is: $P(\pi, s) = \pi(1 - F_q(s))\chi_q - (1 - \pi)(1 - F_u(s))\chi_u$. In the absence of regulation the problem of the recruiter is:

$$\max_{s^A, s^B} \lambda P(\pi^A, s^A) + (1 - \lambda) P(\pi^B, s^B)$$

The first order condition for s^i, $i = A, B$ then becomes:

$$\frac{1 - \pi^i}{\pi^i} g(\hat{s}^i) = r$$

where $g(\theta) \equiv f_u(\theta)/f_q(\theta)$, and $r \equiv \chi_q/\chi_u$. Since $g(\cdot)$ satisfies the MLRP the solution is unique.

Now consider the case where a regulator requires invitation rates to be the same for the two groups. Given belief π and a standard s the fraction of applicants invited for interview is:

$$\rho(\pi, s) \quad = \pi(1 - F_q(s)) + (1 - \pi)(1 - F_u(s))$$

It turns out to be more instructive to pose the problem faced by the recruiter as cost-minimization problem. The cost of choosing standard $\hat{s} + \delta$, where we refer to δ as the *distortion*, can be written as:

$$C(\delta) = \int_{\hat{s}}^{\hat{s}+\delta} \left[r - \frac{1-\pi}{\pi} g(\theta) \right] \mathrm{d}\theta$$

A first order taylor approximation to $g(\theta)$ around $\hat{s}$: $\tilde{g}(\theta) \approx g(\hat{s}) + (\theta - \hat{s})g'(\hat{s})$ allow us to write:

$$\begin{aligned} C(\delta) &= \int_{\hat{s}}^{\hat{s}+\delta} \left[r - \frac{1-\pi}{\pi} g(\theta) \right] \mathrm{d}\theta \\ &= \int_{\hat{s}}^{\hat{s}+\delta} \left[r - \frac{1-\pi}{\pi} (g(\hat{s}) + (\theta - \hat{s})g'(\hat{s})) \right] \mathrm{d}\theta \\ &= \int_{\hat{s}}^{\hat{s}+\delta} \left[\frac{1-\pi}{\pi} (\hat{s} - \theta)g'(\hat{s})) \right] \mathrm{d}\theta \end{aligned}$$

Suppose that $\pi^A > \pi^B$ such that in the unregulated solution the standards used for B's are higher and accordingly invitation rates are lower. Now consider the regulated solution which requires that invitation rates are identical across groups.

The recruiters problem can then be written:

$$\min_{\delta^A, \delta^B} \lambda \int_{\hat{s}^A}^{\hat{s}^A+\delta^A} \left[\frac{1-\pi^A}{\pi^A} (\hat{s}^A - \theta)g'(\hat{s}^A) \right] \mathrm{d}\theta + (1-\lambda) \int_{\hat{s}^B-\delta^B}^{\hat{s}^B} \left[\frac{1-\pi^B}{\pi^B} (\theta - \hat{s}^B)g'(\hat{s}^B) \right] \mathrm{d}\theta$$

$$\text{subject to} \qquad \rho(\pi^A, \hat{s}^A + \delta^A) = \rho(\pi^B, \hat{s}^B - \delta^B)$$

where $\rho(\pi, s) = \int_s^1 h(\pi, \theta)\, \mathrm{d}\theta$ is the invitation rate. We write $h(\pi, \theta) = \pi f_q(\theta) + (1-\pi) f_u(\theta)$ for the signal density.

The first order condition wrt δ^A can be written:

$$\begin{aligned}
-\lambda\frac{1-\pi^A}{\pi^A}\delta^A g'(\hat{s}^A) + \gamma h(\pi^A, \hat{s}^A + \delta^A) = 0 \Rightarrow \\
-\lambda\frac{1-\pi^A}{\pi^A}\delta^A g'(\hat{s}^A) + \gamma(h(\pi^A, \hat{s}^A) + \delta^A h'(\hat{s}^A)) = 0 \Rightarrow \\
\delta^A = \frac{-\frac{\gamma}{\lambda}h(\hat{s}^A)}{-\frac{1-\pi^A}{\pi^A}g'(\hat{s}^A) + \frac{\gamma}{\lambda}h'(\hat{s}^A)}
\end{aligned}$$

where $\gamma < 0$ is the lagrange multiplier. The first line follows from an application of the fundamental theorem of calculus (part I) and the second line relies on a first order taylor approximation to $h(\theta)$ around $\hat{s}^A$: $\tilde{h}(\theta) = h(\hat{s}^A) + (\theta - \hat{s}^A)h'(\hat{s}^A)$. The corresponding first order condition wrt δ^B is:

$$\delta^B = \frac{-\frac{\gamma}{1-\lambda}h(\hat{s}^B)}{-\frac{1-\pi^B}{\pi^B}g'(\hat{s}^B) - \frac{\gamma}{1-\lambda}h'(\hat{s}^B)}$$

The smaller the share of B's in the population the larger the distortion to the unregulated solution for B's. This is intuitive since a large distortion for B's is less costly in terms of foregone profits if they are in a small minority. Also the larger the density the larger the distortion. This is intuitive since the larger the local density the larger the effect on invitation rates.

The one tier insights are easily generalized to two job tiers: L and H, where L and H tier jobs are share μ_L and μ_H respectively, $\mu_L + \mu_H = 1$. Assume $r^H < r^L$ implying tougher standards at the high tier. The problem facing the recruiter can then be written:

$$\begin{aligned}
\min_{\delta_L^A,\delta_H^A,\delta_L^B,\delta_H^B} \lambda \sum_{j=L,H} \mu_j \int_{\hat{s}_j^A}^{\hat{s}_j^A+\delta_j^A} \left[\frac{1-\pi^A}{\pi^A}(\hat{s}_j^A - \theta)g'(\hat{s}_j^A)\right] \mathrm{d}\theta \\
+ (1-\lambda) \sum_{j=L,H} \mu_j \int_{\hat{s}_j^B-\delta_j^B}^{\hat{s}_j^B} \left[\frac{1-\pi^B}{\pi^B}(\theta - \hat{s}_j^B)g'(\hat{s}_j^B)\right] \mathrm{d}\theta \\
\text{subject to} \qquad \sum_{j=L,H} \mu_j \rho(\pi^A, \hat{s}_j^A + \delta_j^A) = \sum_{j=L,H} \mu_j \rho(\pi^B, \hat{s}_j^B - \delta_j^B)
\end{aligned}$$

The first order condition wrt δ_j^A and δ_j^B, $j = L, H$ becomes:

$$\begin{aligned}
\delta_j^A &= \frac{-\frac{\gamma}{\lambda}h(\hat{s}_j^A)}{-\frac{1-\pi^A}{\pi^A}g'(\hat{s}_j^A) + \frac{\gamma}{\lambda}h'(\hat{s}_j{}^A)} \\
\delta_j^B &= \frac{-\frac{\gamma}{1-\lambda}h(\hat{s}_j^B)}{-\frac{1-\pi^B}{\pi^B}g'(\hat{s}_j^B) - \frac{\gamma}{1-\lambda}h'(\hat{s}_j^B)}
\end{aligned}$$

The ratio of the distortions for group B can then be written:

$$\frac{\delta_L^B}{\delta_H^B} = \frac{h(\hat{s}_L^B)}{h(\hat{s}_H^B)} \frac{-\frac{1-\pi^B}{\pi^B} g'(\hat{s}_H^B) - \frac{\gamma}{1-\lambda} h'(\hat{s}_H^B)}{-\frac{1-\pi^B}{\pi^B} g'(\hat{s}_L^B) - \frac{\gamma}{1-\lambda} h'(\hat{s}_L^B)}$$

Thus if group B is relatively more populous around the lower tier standard than at the higher tier standard the distortion at the lower tier will be larger. This would be the case for instance if $h(\theta)$ is monotonically decreasing.

Example - Two Tiers, Linear Signaling Technology Consider the following signaling technology. Signals are distributed on the $[0,1]$ interval. A qualified candidate draws from $f_q(\theta) = 2\theta$, and unqualified candidates draw from $f_u(\theta) = 2(1-\theta)$. Prior beliefs are $(\pi^A, \pi^B) = (.3, .1)$, and job characteristics are fixed at $(r^L, r^H) = (2/1, 3/6)$. Finally $(\mu_L, \mu_H) = (.8, .2)$. Figure 6 illustrates the unconstrained solution.

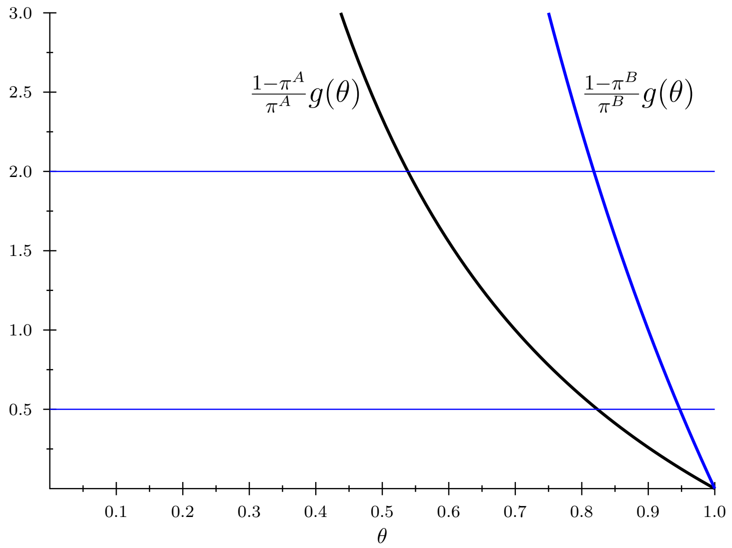
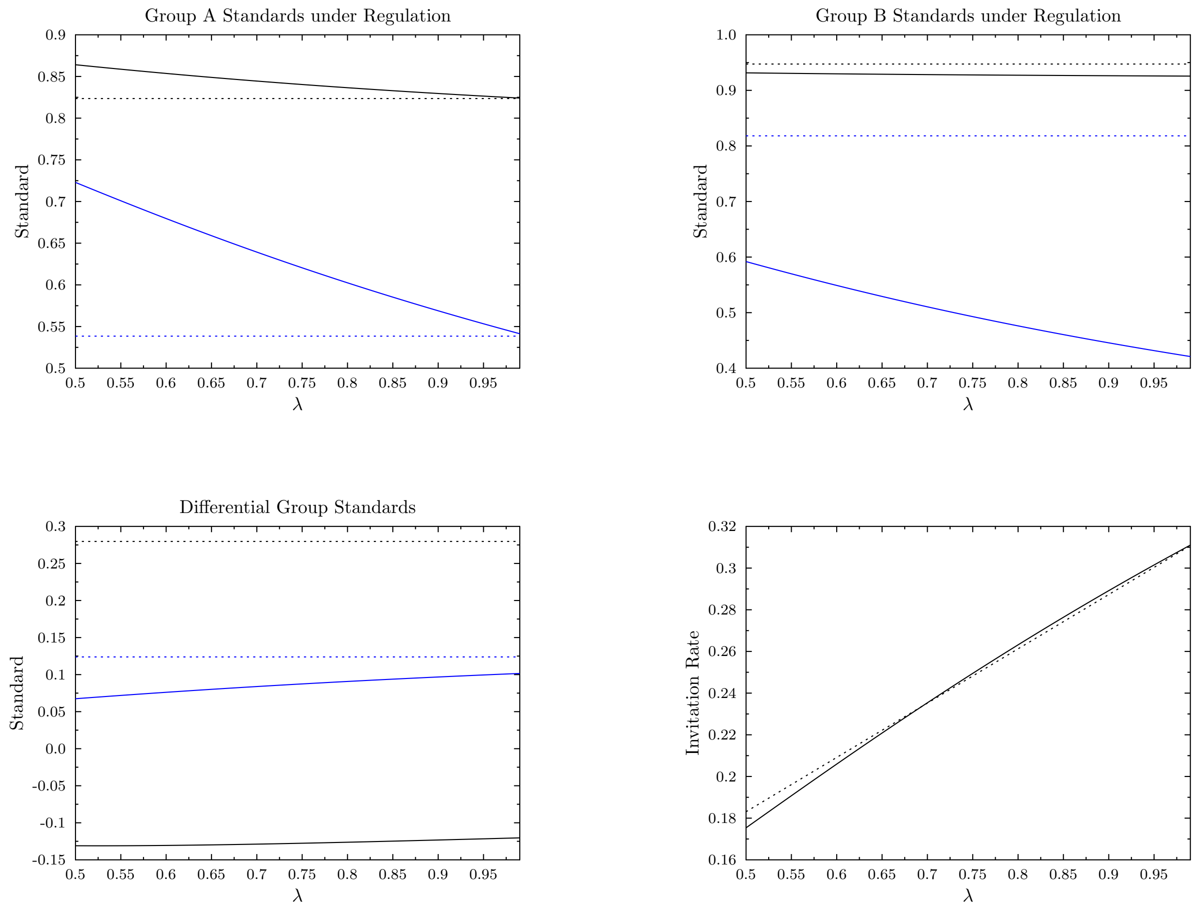


Figure 6: Unconstrained optimal standards. $(\pi^A, \pi^B) = (.3, .1)$, and $(r^L, r^H) = (2/1, 3/6)$.

The constrained solution requires that the recruiter achieves equal invitation rates across groups. Figure 7 illustrates the solution to this problem as a function of the majority share of the population. The two top panels are for group A and B respectively. The bottom panel illustrates the population wide invitation rate under the constraint.



Notes: $(\pi^A, \pi^B) = (.3, .1)$, $(r^L, r^H) = (2/1, 3/6)$, $(\mu_L, \mu_H) = (.8, .2)$. Top Panels: Optimal standards under regulation by group. Solutions to unconstrained problem in dashed pen. Standard for lower tier job in blue. Lower Left Panel: Group Differential in Standards (Standard group B - Standard group A) by tier. High tier in blue pen. Unconstrained differential in dashed pen. Lower Right Panel: Population wide promotion rates. Unconstrained promotion rate in dashed pen.

Figure 7: Optimal Standards and Promotion Rates under Regulation.

B Supplementary Regressions

		Control				**Treatment**			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Majority		0.0744**	0.0765**	0.0819**	0.0917***	0.00928	0.0191	0.0180	0.0290
		(0.0337)	(0.0329)	(0.0358)	(0.0336)	(0.0317)	(0.0318)	(0.0330)	(0.0339)
Female				0.0225	0.0682**			0.00940	0.0117
				(0.0342)	(0.0343)			(0.0282)	(0.0300)
Age	30-45			-0.00236	0.0183			0.0456	0.0429
				(0.0370)	(0.0361)			(0.0310)	(0.0305)
	45-65			-0.0294	-0.00381			-0.0579*	-0.0586*
				(0.0444)	(0.0430)			(0.0295)	(0.0339)
Observations		511	511	511	511	638	638	638	638
R-squared		0.007	0.138	0.070	0.184	0.000	0.088	0.048	0.125
Job FE		NO	YES	NO	YES	NO	YES	NO	YES
Matchspecific controls		NO	NO	YES	YES	NO	NO	YES	YES

Notes: Robust standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.1

Table 13: Interview Decision: Separate Regressions

		Control				Treatment			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Majority		-0.0303	-0.0771	0.00390	-0.0154	-0.0185	-0.0795	-0.0530	-0.0658
		(0.190)	(0.187)	(0.161)	(0.210)	(0.172)	(0.222)	(0.181)	(0.274)
Female				0.0360	0.105			-0.0948	-0.171
				(0.141)	(0.173)			(0.174)	(0.228)
Age	30-45			-0.216	-0.280			0.137	0.0888
				(0.165)	(0.181)			(0.190)	(0.252)
	45-65			-0.295	-0.344			-0.0681	-0.0564
				(0.198)	(0.239)			(0.233)	(0.306)
Observations		74	74	74	74	63	63	63	63
R-squared		0.000	0.313	0.192	0.390	0.000	0.147	0.083	0.212
Job FE		NO	YES	NO	YES	NO	YES	NO	YES
Matchspecific Controls		NO	NO	YES	YES	NO	NO	YES	YES

Notes: Robust standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.1

Table 14: Offer Decision: Separate Regressions. Interview Sample.

		Control				Treatment			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Majority		0.0324	0.0286	0.0462*	0.0425*	0.00124	0.000605	0.00525	0.00390
		(0.0243)	(0.0247)	(0.0254)	(0.0258)	(0.0188)	(0.0168)	(0.0190)	(0.0168)
Female				0.0156	0.0386			-0.00805	-0.0142
				(0.0243)	(0.0275)			(0.0153)	(0.0156)
Age	30-45			-0.0223	-0.0177			0.0309*	0.0292*
				(0.0294)	(0.0288)			(0.0185)	(0.0170)
	45-65			-0.0340	-0.0321			-0.0179	-0.0226
				(0.0337)	(0.0339)			(0.0153)	(0.0176)
Observations		511	511	511	511	638	638	638	638
R-squared		0.002	0.086	0.046	0.113	0.000	0.059	0.030	0.083
Job FE		NO	YES	NO	YES	NO	YES	NO	YES
Matchspecific Controls		NO	NO	YES	YES	NO	NO	YES	YES

Notes: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 15: Offer Decision: Separate Regressions. Full Sample.

	Offer Decision			
	(1)	(2)	(3)	(4)
Majority	-0.0303	-0.0771	0.0149	-0.0491
	(0.190)	(0.187)	(0.156)	(0.194)
Majority×Treatment	0.0118	-0.00238	-0.0375	-0.0390
	(0.256)	(0.290)	(0.232)	(0.299)
Female			-0.0130	0.0211
			(0.0996)	(0.130)
Age 30-45			-0.0662	-0.0900
			(0.119)	(0.153)
45-65			-0.183	-0.214
			(0.138)	(0.188)
Observations	137	137	137	137
R-squared	0.025	0.261	0.092	0.304
Job FE	NO	YES	NO	YES
Matchspecific Controls	NO	NO	YES	YES
$\delta + \tau_1 = 0$ (p-value)	0.914	0.721	0.898	0.706

Notes: Robust standard errors in parentheses. *** p <0.01, ** p <0.05, * p <0.1

Table 16: Offer Decision. Interview Sample.

C Robustness

The choice of the linear probability model for estimation is a choice of convenience. This raises the question whether our estimates are robust to more standard approaches to discrete choice. In this section we show that our results are essentially robust to such alternative specifications. We present estimates from two different specifications. The first specification is standard probit without job level fixed effects. In the second specification we introduce job level fixed effects. No known consistent estimator exists in the latter case (Greene, 2001)[34]. This problems is the more serious the fewer applicants there are.

[34]Unlike the linear case de-meaning does not get rid of the fixed effects in the likelihood function. This method thus suffers from the incidental parameter problem since the number of parameters increase with the number of jobs.

Simulations reported in Greene (2001) suggest however that the problem is less severe for $T_j > 10$, and so for this specific application the problem seems to be less severe. Approaches exists that allow consistent estimates of the coefficients, but since these methods involves "averaging" out the fixed effects, the fixed effects are not estimated and we can accordingly not calculate marginal effects or average partial effects which are of great interest here. Alternatively we could opt for a random effects specification. Such estimates are only consistent if the effect is not correlated with the explanatory variables, something which does not seem plausible in our setting.

		Invited for Interview				
Model		LPM (+FE)	Probit		Probit FE	
		ME	ME	APE	ME	APE
Majority		0.0873***	0.0663***	0.0808***	0.0635***	0.0849***
		(0.0300)	(0.0209)	(0.0308)	(0.0164)	(0.0290)
Majority×Treatment		-0.0556	-0.0661*	-0.0678*	-0.0556	-0.0593
		(0.0398)	(0.0395)	(0.0409)	(0.0383)	(0.0413)
Female		0.0398	0.0229	0.0236	0.0367*	0.0399*
		(0.0250)	(0.0206)	(0.0216)	(0.0199)	(0.0223)
Age	30-45	0.0328	0.0235	0.0236	0.0291	0.0303
		(0.0244)	(0.0247)	(0.0247)	(0.0232)	(0.0239)
	45-65	-0.0314	-0.0458	-0.0493	-0.0276	-0.0305
		(0.0319)	(0.0319)	(0.0364)	(0.0304)	(0.0353)
Education	Middle	0.0316	0.0473	0.0477	0.0413	0.0433
		(0.0247)	(0.0378)	(0.0381)	(0.0310)	(0.0325)
	High	-0.00496	0.0468	0.0462	0.0149	0.0155
		(0.0300)	(0.0437)	(0.0420)	(0.0351)	(0.0363)
Observations		1,148	1,148	1,148	1,148	1,148
R-squared		0.146	-	-	-	-
Job FE		YES	NO	NO	YES	YES
$\ln L_0$		-	-419.8	-419.8	-417.7	-417.7
$\ln L$		-	-393.2	-393.2	-346.3	-346.3

Notes: Clustered standard errors (per vacancy) in parentheses. ME: Marginal Effects, APE: Average Partial Effects, *** p <0.01, ** p <0.05, * p <0.1

Table 17: Interview Decision. Full Sample.

		Job Offer				
Model		LPM (+FE)	Probit		Probit FE	
		ME	ME	APE	ME	APE
Majority		0.0365	0.0270**	0.0385*	0.0189*	0.0308
		(0.0239)	(0.0113)	(0.0200)	(0.00986)	(0.0204)
Majority×Treatment		-0.0279	-0.0289	-0.0326	-0.0186	-0.0244
		(0.0398)	(0.0239)	(0.0269)	(0.0220)	(0.0288)
Female		0.0123	0.00804	0.00915	0.0105	0.0141
		(0.0176)	(0.0125)	(0.0145)	(0.0109)	(0.0150)
Age	30-45	0.00778	0.00838	0.00928	0.00465	0.00602
		(0.0147)	(0.0126)	(0.0139)	(0.0114)	(0.0147)
	45-65	-0.0260	-0.0237	-0.0293	-0.0201*	-0.0294
		(0.0181)	(0.0147)	(0.0199)	(0.0121)	(0.0200)
Education	Middle	0.0233*	0.0208	0.0230	0.0262*	0.0333*
		(0.0134)	(0.0197)	(0.0216)	(0.0151)	(0.0187)
	High	-0.0137	0.00514	0.00570	0.0109	0.0137
		(0.0175)	(0.0210)	(0.0231)	(0.0177)	(0.0218)
Observations		1,148	1,148	1,148	1,148	1,148
R-squared		0.086	-	-	-	-
Job FE		YES	NO	NO	YES	YES
Clusters		36	36	36	36	36
$\ln L_0$		-	-217.8	-217.8	-217.8	-217.8
$\ln L$		-	-200.1	-200.1	-182.3	-182.3

Notes: Clustered standard errors (on job level) in parentheses. ME: Marginal Effects, APE: Average Partial Effects, *** p <0.01, ** p <0.05, * p <0.1

Table 18: Offer Decision. Full Sample.

		Job Offer (Conditional on Interview)		
Model		LPM (+FE)	Probit	
		ME	ME	APE
Majority		-0.0491	-0.00116	-0.00108
		(0.194)	(0.173)	(0.161)
Majority×Treatment		-0.0390	-0.0287	-0.0269
		(0.298)	(0.251)	(0.236)
Female		0.0211	0.0262	0.0246
		(0.130)	(0.0999)	(0.0937)
Age	30-45	-0.0900	-0.0336	-0.0313
		(0.152)	(0.119)	(0.111)
	45-65	-0.214	-0.160	-0.157
		(0.188)	(0.126)	(0.131)
Education	Middle	0.187	0.0243	0.0227
		(0.192)	(0.204)	(0.191)
	High	-0.101	-0.163	-0.154
		(0.309)	(0.225)	(0.215)
Observations		136	136	136
R-squared		0.297	-	-
Job FE		YES	NO	NO
$\ln L_0$		-	-91.36	-91.36
$\ln L$		-	-85.53	-85.53

Notes: Robust standard errors in parentheses. ME: Marginal Effects, APE: Average Partial Effects, *** p <0.01, ** p <0.05, * p <0.1

Table 19: Offer Decision. Interview Sample.

D The Follow Up Experiment

D.1 Descriptive Statistics

Required	Number of Vacancies			Number of Applicants		
Education	Treatment	Control	Total	Treatment	Control	Total
Low	7	10	17	237	421	658
Medium	13	6	19	319	163	482
High	4	8	12	81	172	253
Total	24	24	48	637	756	1393

Table 20: Vacancies and Applicants by Required Education Level (Follow Up - Full Sample)

Required Education		**Majority**	**Female**	**Age** <30	30-45	45-65	**Education** Low	Medium	High	**N**
Low	Control	.77	.587	.257	.385	.333	.23	.511	.211***	421
		(.422)	(.493)	(.437)	(.487)	(.472)	(.422)	(.5)	(.409)	
	Treatment	.781	.62	.249	.384	.338	.0506	.738	.198	237
		(.415)	(.486)	(.433)	(.487)	(.474)	(.22)	(.44)	(.4)	
Middle	Control	.92***	.313***	.245	.503	.252	.0245	.245	.724	163
		(.272)	(.465)	(.432)	(.502)	(.435)	(.155)	(.432)	(.448)	
	Treatment	.809	.632	.229	.42	.307	.0063	.295	.683	319
		(.394)	(.483)	(.421)	(.494)	(.462)	(.0791)	(.457)	(.466)	
High	Control	.959**	.471	.105	.547	.32	.0116	.0814	.89	172
		(.198)	(.501)	(.307)	(.499)	(.468)	(.108)	(.274)	(.314)	
	Treatment	.877	.412	.0864	.506	.358	0	.0617	.938	81
		(.331)	(.495)	(.283)	(.503)	(.482)	(0)	(.242)	(.242)	
Total	Control	.845*	.501***	.22	.447	.312	.136	.356	.476***	756
		(.362)	(.5)	(.414)	(.498)	(.464)	(.343)	(.479)	(.5)	
	Treatment	.807	.6	.218	.418	.325	.022	.43	.535	637
		(.395)	(.49)	(.413)	(.494)	(.469)	(.147)	(.495)	(.499)	

Notes: Variable means. Standard deviations in parenthesis. Mann-Whitney ranksum means test for majority and gender variables. Pearson's test for age and education categories. *** p <0.01, ** p <0.05, * p <0.1. 41 observations has missing age category. 26 observation has missing education category.

Table 21: Applicant Descriptive Statistics by Required Education Level (Means). Full Follow Up Sample.

Group		Invited	Offer	Female	Age			Education			N
					<30	30-45	45-65	Low	Medium	High	
Minority	Control	.162	.0427	.59**	.359	.453	.162	.103	.504	.333**	117
		(.37)	(.203)	(.494)	(.482)	(.5)	(.37)	(.305)	(.502)	(.473)	
	Treatment	.203	.0488	.746	.293	.439	.187	.0244	.553	.407	123
		(.404)	(.216)	(.437)	(.457)	(.498)	(.391)	(.155)	(.499)	(.499)	
Majority	Control	.177**	.0376	.485***	.194	.446	.34	.142	.329	.502***	639
		(.382)	(.19)	(.5)	(.396)	(.497)	(.474)	(.35)	(.47)	(.5)	
	Treatment	.224	.0486	.565	.2	.412	.358	.0214	.401	.566	514
		(.417)	(.215)	(.496)	(.401)	(.493)	(.48)	(.145)	(.491)	(.496)	

Notes: Variable means. Standard deviations in parenthesis. Mann-Whitney ranksum means test for invited, offer and gender variables. Pearson's test for age and education categories. *** p <0.01, ** p <0.05, * p <0.1. 41 observations has missing age category. 26 observation has missing education category.

Table 22: Applicant Descriptive Statistics by Group Membership (Means). Full Sample.