Stereotypes and Willingness to Change Them: Testing Theories of Discrimination in South Africa

Jorge M. Agüero*

June, 2008

Abstract

Employers often decide job assignments or wages after observing *productivity* signals from workers. Discrimination can occur because employers have stereotypes (priors) against a group of workers, or because they use signals differently depending on the worker's group. This paper introduces an estimable Bayesian framework that allows us to recover both the priors and the updating behavior of *evaluators* who observe noisy signals from *candidates*. Using data from a quasi-experiment in South Africa I test for the precise form of racial discrimination. I find evidence of discrimination without overtly negative priors. Discrimination occurs because white evaluators use signals to update their priors about white candidates but not when evaluating black candidates. Blacks, on the other hand, use signals to update their priors about all candidates. The paper uses the estimated structural parameters to simulate how evaluators would choose among equally performing candidates as a tool to show the relative importance of stereotypes and updating behavior on discrimination.

Keywords: Discrimination, Experiments, Games, Bayesian Learning, South Africa.
JEL codes: C9, J15, J71, C11, O5.

---

*Department of Economics, University of California, Riverside, 4108 Sproul Hall, Riverside CA 92521; email:jorge.aguero@ucr.edu. I would like to thank Michael Carter, James Walker and Maurizio Mazzocco for their support and continuous comments and suggestions. I also benefited from conversations and comments from seminar participants at the University of Wisconsin-Madison, the University of KwaZulu-Natal, the University of Cape Town, the Group of Analysis for Development (GRADE) and the Northeast Universities Conference on Development at Brown University. Chantal and Crystal Munthree and Ingrid Woolard helped with the collection of the raw data. Nozipho Ntuli, Thulani Gwala, Thabani Buthelezi and Mimi Ndokweni provided insights about the identification of race and other characteristics of the game participants. Duncan Irvine and Kee-Leen Irvine, from Rapid Blue, answered my questions about the game. Michele Back helped me edit this manuscript, however, all remaining errors are my own.

"Not to know is bad. Not to wish to know is worse." -Wolof proverb.

1 Introduction

A recent paper by Bertrand and Mullainathan (2004) finds important evidence of racial discrimination in the US labor market. The authors sent fictitious resumes in response to want ads, where the resumes were the same except for the name of the applicants. Resumes with "white names" (e.g., Emily and Greg) received 50% more callbacks than resumes with "black names" (e.g., Lakisha and Jamal). But why did Lakisha's resume generate few callbacks? One explanation is that employers begin with such a low prior for her skills that even good credentials could not put her over the callback threshold. On a somewhat deeper and more pernicious level, employers could be unable (or unwilling) to see beyond Lakisha's race and update their evaluation based on the information contained on her resume.[1] The goal of this paper is to devise a Bayesian framework and use data from a quasi-experiment in South Africa to identify the relative importance of priors and updating behavior (i.e., the willingness to change priors) on discrimination.

I model a situation where evaluators have to decide whether candidates are capable of performing a task. Evaluators cannot observe candidates' abilities or qualifications for the task. They only observe each candidate's group and a noisy signal about the candidate's ability. Evaluators use their priors and the signals to form their posterior beliefs using Bayes rule.

Modeling evaluators as Bayesian agents allows us to break down their decision process in two parts. First, evaluators use candidates' observable characteristics to infer their ability. In this model, priors serve as the stereotypes.[2] These priors might not be accurate, implying negative consequences for a group of candidates. This is the analog for the "not to know" portion of this paper's opening quote. Second, Bayesian agents update their priors after observing information in order to form the posterior. However, it might be the case that evaluators refuse to update their priors after seeing the signals if they consider them "uninformative." This paper exploits this idea in order to test whether agents use information

[1] The authors' complementary finding that credentials have a positive but lower return for blacks only confirms that signals are taken into account by employers.

[2] In this paper I use the social psychology definition of "stereotypes" as attaching (removing) a characteristic to (or from) a person because he or she belongs to a certain group (Banaji 2002).

(such as the credentials on Lakisha's resume) in the same way for all candidates. "Not wishing to know" occurs when evaluators do not update their priors for a group of candidates. By decomposing the evaluator's decision process I can identify the sources of discrimination.

Understanding the sources of discriminatory behavior is key for the design of antidiscrimination policies, such as those intended with affirmative action.[3] This understanding also motivates studies explaining the persistent difference in earnings between racial or ethnic groups in labor markets (Altonji and Blank 1999). Similarly, in many developing countries the rich and the poor differ in more than just asset holdings; they are also of different races (e.g Psacharopoulos and Patrinos 1994, Carter and May 2001). Therefore, knowing how discrimination operates might help us explain the persistence of poverty. Despite wide interest in the topic, when evidence of discrimination is found, the economic literature is basically silent about the causes of discrimination (e.g. Ayres and Siegelman 1995, Goldin and Rouse 2000, Neumark 1996, Bertrand and Mullainathan 2004).

To fill this gap, I introduce a Bayesian framework that uses aspects from two sets of models explaining discrimination. The first set explains discrimination as the existence of negative stereotypes about the capabilities of certain groups of the population (e.g., Arrow 1973, Phelps 1972, Coate and Loury 1993). In such models, priors are updated in the same way for all workers. On the other hand, the work by Aigner and Cain (1977), Lundberg and Startz (1983) and Lundberg (1991) depart from the framework of negative stereotypes by assuming that employers have the same prior beliefs for all workers. Here, candidates are seen by employers as *ex ante* identical across groups, but the signal is modeled as less informative for a certain group of candidates.[4]

While this second set of papers provides an alternative explanation for discrimination, studies in social psychology suggest that stereotypes are inevitable, immediate and intrinsic to the process of perceiving (e.g. Banaji 2002, Fiske 1998). Hence, it would be inadequate to rule out the possibility of negative stereotypes. By incorporating both arguments –priors and and differential treatment of the observed signals– this paper offers a broader set of explanations about the sources of discrimination.

The proposed framework shares some features with the models described above but there

[3]See Coate and Loury (1993) for a discussion on whether affirmative action policies can achieve this goal.

[4]Lang (1986) argues that language and culture could explain this feature. For example, a white male manager would have more difficulty evaluating female or black workers. See also Altonji and Blank (1999) for a review.

are four important differences. First, I use the results from research in social psychology, where the accuracy of the evaluators' belief is no longer relevant. Unlike Coate and Loury (1993), the introduced model does not require stereotypes to correctly describe the essence of the group (Banaji 2002, p. 15101). Second, the model does not take into account educational or any other human capital investments made by candidates, focusing only on the decisions made by evaluators.

Third, evaluators are not required to update priors about all candidates in the same way (as long as they follow Bayes rule). As in Lundberg and Startz (1983) and others, the signals can be treated in different ways for different groups of candidates. Hence, in the evaluator's mind, the probability of observing a good signal depends not only on the ability of the candidate (i.e., the likelihood ratio in a Bayesian framework,) but also on the candidate's group identity. Fourth, discrimination in this model can occur not only because of differences in prior beliefs about groups, but also due to differences in the updating parameters regarding the candidate's group.

A natural way to test the model is to have an experiment where strangers have to guess at an unobserved measure of ability and reveal their posteriors after observing signals related to the unobserved ability. Such an experiment exists in the form of television game show *The Weakest Link*. In this show nine strangers compete for a winner-take-all prize by answering trivia questions. The prize increases with the number of correct answers. The players' performance is a noisy signal of their ability, because the difficulty of the question is random and uncorrelated with the players' performance. Players vote off one contestant at the end of each round. I assume that expected income maximizing players (motivated by the high stakes of the game) would vote against the player they believe is the one with the lowest ability, at least during the first round of the game. Because players do not observe each other's ability but only physical characteristics such as race, the formation of stereotypes is highly possible. Under the assumption of voting against the weakest player, the voting behavior is a (discrete) realization of the posterior.

This paper uses data from the South African version of the show to test for racial discrimination. Using a behavioral model I can identify the parameters that would allow me to test whether negative stereotypes exist and whether people update in the same way for black and white players.[5]

[5]Two other papers use the US version of the show to distinguish among a different set of theories of

Some caveats apply when such a dataset is used. They are described in more detail in section 4.1 but summarized as follows. First, the fact that the show is broadcast on national television might preclude discriminatory behavior, biasing our results toward no discrimination. Second, the sample is not a random draw of the population and has an urban bias. Whites are overrepresented and players are highly educated compared to the population figures. However, the selection process created a sample where blacks and whites do not vary by other observable characteristics. While this does not allow us to expand the results to the entire population of South Africa, it does allow us to isolate the role of race in the game. Hence, for the purpose of the paper, the sample selection is a plus rather than a drawback.

Using reduced form estimates I show that player performance is a good "predictor" of voting behavior: the worse they play the game, the more votes they receive. However, the number of votes received has a racial bias even after controlling for performance. This is the analog for the Emily and Lakisha problem stated above, meaning that the "candidate's" race remains important even after controlling for "credentials". Having a Bayesian model such as the one described above allows us to separate the observed discrimination into priors and willingness to change them, thus providing a better insight of the nature of discrimination.

The main result of the paper shows no evidence of negative priors against either group. However, this does not preclude discrimination. White players behave as if they refuse to update their priors about blacks, but they are willing to do so for other white participants. They treat all black candidates the same, regardless of their performance. In contrast, blacks update their priors for both races.

The rest of the paper is divided in seven sections. Section two briefly reviews previous measures of discrimination. Section three presents the model and its testable implications. The data and the estimation strategy are described in sections four and five. Section six discusses how performance and voting patterns relate to people's race and the main results are shown in section seven. This section includes robustness checks and a simulation showing

discrimination: preferences (Becker 1957) and information (Arrow 1973, among others). Both papers agree with the assumption that during the initial rounds of the game players will find it optimal to vote against the weak player. Both papers also use the dynamics of the game to distinguish between their theories of discrimination, but these dynamics might be affected by issues such as reputation, vengeance and disclosure of information. Using a model of Bayesian learning I can estimate the priors and how they are updated by using only data from the first round where there is no history, thus avoiding the problems from the existing literature.

the relative importance of priors and the way different evaluators update (or do not) their priors on discrimination. Section eight summarizes the paper and discusses the limitations and pending issues.

2 Previous measures of discrimination

Measuring discrimination is a difficult task. During the apartheid era in South Africa and before the Civil Rights movement in the United States, there were laws that separated groups of the population. The discourse in the employment ads during those times shows clear evidence of discrimination (Darity and Mason 1998). The current absence of these events is an improvement, but discrimination continues in more subtle ways.

In economics, a common approach to measure discrimination is to decompose differences in wages (or in labor force participation) for two groups into observed and unobserved factors using the Oaxaca-Blinder decomposition. The observed factors include schooling and experience in the labor force as well as the returns of these variables. The unobserved factors are used as a proxy for discrimination. This methodology has been used in both developing and developed countries[6] and has the advantage of using household-level datasets, allowing researchers to draw conclusions about the population. However, this approach has been criticized as an inadequate approximation for discrimination, as discrimination can also affect observed factors such as schooling and experience in the labor force (Altonji and Blank 1999). Thus, the unobserved differences might not capture the full extent of discrimination.

Several alternatives have been explored to avoid this problem by using data from less conventional sources. The goal of these alternatives is to find clearer evidence of discrimination[7]. For example, Ayres and Siegelman (1995) created audits where trained individuals from different races and genders bargained for a new car. The authors' findings suggest that dealers quoted lower prices for whites than blacks or female buyers using identical scripted bargaining strategies. Goldin and Rouse (2000) evaluate the impact of "blind" auditions on hiring female musicians in orchestras. They found that females have a much higher probability of moving to higher rounds of the auditions when performing behind a screen. The

[6]See for example Altonji and Blank (1999) for applications in the U.S. and Lam and Leibbrandt (2004) and Casale (2003) for examples about South Africa.

[7]See Anderson, Fryer, and Holt (2005) for a survey of experiments measuring discrimination.

work by Bertrand and Mullainathan (2004) described in the introduction also falls into this category.

Three conclusions can be drawn from the literature searching for evidence of discrimination. First, in order to test for discrimination, scholars are moving away from traditional household-level datasets. The studies mentioned above are closer to case studies and hence cannot make inferences about the entire population. The advantage, though, is to have a clearer way to find evidence of discrimination. Second, unlike the Oaxaca-Blinder approach, the study of discrimination using these new methods has focused mostly on the United States, with almost no evidence from developing countries[8]. Third, all of these studies, including those using the Oaxaca-Blinder approach from developed and developing countries, are mute with respect to cause of discrimination. When evidence of discrimination is found, we do not know the reasons driving this behavior. In the next section I introduce a model that allow us decompose discrimination into differences in priors (stereotypes) and differences in how information is treated (whether or not priors are updated).[9]

3 The model and testable implications

3.1 A model of learning

The model presented here is one of Bayesian learning. The idea is to have a group of evaluators and candidates who do not know each other, where the former have to choose which of the latter is most likely to not qualify for a task. Evaluators make decisions about the candidates' unobservable characteristics (such as productivity or ability). To do this, evaluators approximate the unobservable characteristics with observable characteristics (such as race.) Evaluators will also observe a "noisy signal" that is imperfectly related to the candidates' productivity.

The Bayesian part of the model comes from the assumption about how candidates learn. First, for each observable characteristic –in this paper, race– evaluators have a prior belief

[8]Moreno, Ñopo, Saavedra, and Torero (2004) provide preliminary evidence of audit studies in Peru. See also Frijters (1999) for South Africa.

[9]In the appendix we discuss how the methodology developed here could also be applied to distinguish between discrimination based on preferences (Becker 1957) and discrimination based on information (e.g., Arrow 1973, Phelps 1972).

about the proportion of black candidates with "high" or "low" productivity, and a corresponding belief for white candidates. Second, evaluators have a probability distribution for the likelihood to observe a "good" signal from a black candidate with a low (or high) productivity, and a similar distribution for whites. In Arrow's (1973) and Coate and Loury's (1993) models, the likelihoods do not vary by the race of the candidate. In this model, all beliefs and likelihoods are predetermined and embedded in the evaluator's minds before meeting the candidates. Then information is revealed in the form of signals. Each evaluator observes the (noisy) signals from each candidates (e.g. the results of a test, or how well candidates answer a set of questions). Using these signals, together with the priors and the likelihoods, each evaluator constructs his/her posterior belief about the probability that a candidate has a low or high ability following Bayes theorem.

Formally, let i index the evaluators who belong to a class $\mathcal{E}$. Candidates will be indexed by k. They can belong to different groups indexed by j, so $j = \{1, \ldots, J\}$. The key idea for the identification of the parameters of interest is that all evaluators treat candidates in group j identically. I also assume that all evaluators in class $\mathcal{E}$ behave in the same manner. In other words, we are able to identify only how the average evaluator from class $\mathcal{E}$ deals with the average candidate from group j. I will return to this point later in section 5.

I define j as an observable characteristic of the candidates: in this paper, race. In the case of South Africa, j takes two values; $j = 1$ for blacks (which includes Africans, Coloured (mixed race) and Indians) and $j = 2$ for whites. Let θ represent a candidate's unobservable characteristic and assume that θ is binary: $\theta = 1$ when ability is high and $\theta = 0$ when ability is low. This is the parameter evaluators would like to know about the candidates but do not. I am also assuming that there is no heterogeneity within each value of θ. Let y_{jk} denote the quality of the signal from candidate k that belongs to group j. When the signal is "good" $y_{jk} = 1$, otherwise $y_{jk} = 0$. Unlike Lang (1986) the quality is not decided by the evaluator. We consider the case where –consistent with the experiment used in this paper– an outside "judge" determines whether the signal is good or not. This point will become clearer when I present the data in section 4.1. The quality of the signal, together with the candidate's performance, is public information.

I now turn to the prior beliefs. Because θ is not observable it is natural to think that players have a prior belief or stereotype about the proportion of blacks and whites with low ability. This idea is reinforced by the studies done in social psychology discussed in the

introduction. Let $\text{Prob}_i(\theta_j = 0) = \alpha^0_{\mathcal{E},j}$ be the probability that players from group j have low ability from the point of view of player $i \in \mathcal{E}$. To save on notation and because all $i \in \mathcal{E}$ have the same set of priors about members of group j, let me erase the $\mathcal{E}$-subscript, so I will refer to α^0_j instead.

Here I am assuming that evaluators are "certain" about their priors. For example, when the prior for blacks is said to be 0.5, the evaluator having this prior is not allowed to have uncertainty about it. That is, the person thinks that that the prior has a probability equal to one of being true. This is of course, a strong assumption, but I assume this for three reasons. First, because it reduces the number of parameters (I will need another parameter for the variance, as in the Beta distribution.) Second, as I discuss later, given the fact that the observed posterior (the evaluator's vote) takes the form of a discrete binary variable (I only know which candidates have been rejected by evaluator i) there is limited information about the variance of the posterior. Third, it is common when dealing with multivariate discrete choices to assume a fixed value for the variance. In the case of the logit model, the variance does not depend on the parameters and in the case of the probit the variance is usually assumed to be equal to one. It is important to note that the assumption about the certainty of the priors does not imply that people are not willing to change them. The assumption made is about the variance of the prior, which can also be shaped by the revelation of information. In that sense, this paper evaluates the willingness to change the "mean" of the prior, keeping the variance constant.

Let q^H_j be the likelihood that evaluator i relates a "good" signal ($y_{jk} = 1$) as coming from a high ability candidate from group j, and let q^L_j be the analog for a low ability candidate. Notice that q^H_j and q^L_j do not need to add up to one, that is why they are called *likelihood* parameters and not probabilities. By allowing q^H_j and q^L_j to vary by the race of the candidate I incorporate the idea behind the papers by Aigner and Cain (1977) and Lundberg and Startz (1983), that is, letting the evaluator treat the candidate differently beyond unequal priors. Otherwise, if $q^z_j = q^z_j$ for all $z = \{H, L\}$, we are back in the Coate and Loury (1993) type of models.

Assume that the noisy signal is observed S times, so the number of good signals is given by $Y_{jk} = \sum^S y_{jk}$. The probability of observing Y_{jk} follows the binomial distribution below

when a candidate has high ability:

$$\text{Prob}_i(Y_{jk}|\theta_j = 1) = \frac{S!}{Y_{jk}!(S-Y_{jk})!} q_j^{H\,Y_{jk}} (1-q_j^H)^{S-Y_{jk}} = g_i^H(Y_{jk}) \tag{1}$$

and similarly for a candidate with low ability:

$$\text{Prob}_i(Y_{jk}|\theta_j = 0) = \frac{S!}{Y_{jk}!(S-Y_{jk})!} q_j^{L\,Y_{jk}} (1-q_j^L)^{S-Y_{jk}} = g_i^L(Y_{jk}) \tag{2}$$

The assumption of Bayesian updating defines the way evaluators modify their beliefs after the information is revealed. Let α_{jk}^1 be the posterior probability that evaluator i assigns to the k-th candidate, belonging to group j, as having a low ability after observing k's signals summarized by Y_{jk}. Formally,

$$\begin{aligned}
&\text{Prob}_i(\text{player k} \in j \text{ is low type } |k \text{ has } Y_{jk} \text{ good signals}) \\
&\quad = \text{Prob}_i(\theta_j = 0|Y_{jk}) \\
&\quad = \frac{\text{Prob}_i(Y_{jk}|\theta_j = 0)\text{Prob}_i(\theta_j = 0)}{\text{Prob}_i(Y_{jk}|\theta_j = 0)\text{Prob}_i(\theta_j = 0) + \text{Prob}_i(Y_{jk}|\theta_j = 1)\text{Prob}_i(\theta_j = 1)} \\
&\quad = \frac{g_i^L(Y_{jk})\alpha_j^0}{g_i^L(Y_{jk})\alpha_j^0 + g_i^H(Y_{jk})(1-\alpha_j^0)} \\
&\quad = \alpha_{jk}^1(\alpha_j^0, q_j^H, q_j^L; Y_{jk})
\end{aligned} \tag{3}$$

where $g_i^H(\cdot)$ and $g_i^L(\cdot)$ are functions defined in equations (1) and (2,) respectively. The posterior probability α_{jk}^1 is then a (nonlinear) function of the structural parameters of the model $(\alpha_j^0, q_j^H, q_j^L)$ for all j, as well as the information revealed from each candidate in the form of Y_{jk} for all k and all j.

Note that when $q_j^H = q_j^L$ for some j, equations (2) and (1) are identical. In that case, in equation (3) the functions g_i^H and g_i^L cancel out from the numerator and denominator. What is left in equation (3) is α_j^0 in the numerator and $\alpha_j^0 + 1 - \alpha_j^0$ in the denominator. This in turn implies that the posterior probability α_{jk}^1 is equal to the prior probability α_j^0 for all $k \in j$. I will use this feature to develop the test for the willingness to update priors.

3.2 Testable implications

A test for discrimination can be developed first by observing differences in prior beliefs. If evaluator i believes that a candidate from group j has a higher probability to be a low ability person than a candidate from group t, is evidence in favor of the existence of negative stereotypes against group j and constitutes the first test.

Test 1 (Negative stereotypes) *If $\alpha_j^0 > \alpha_t^0$ for some $t \neq j$ then negative stereotypes exist about members of group j. Otherwise candidates from both races are treated equally, at least initially.*

Test 1 implies that negative stereotypes appear in this model when the only reason to believe that candidate k has a higher probability of being a low ability person is k's race. This is also the definition used in Coate and Loury (1993).

The second aspect of a discriminatory behavior is the one regarding the values of q_j^H and q_j^L. Consider the extreme case described above when $q_j^H = q_j^L$. I now present a way to distinguish tests for the willingness to update beliefs:

Test 2 (Unwillingness to change) *If $q_j^H = q_j^L$, evaluators are not willing to change their initial beliefs about candidates from group j.*

As shown above, if for some j we have $q_j^H = q_j^L$ it implies that $\alpha_j^0 = \alpha_j^1$ in equation (3), that is, prior posterior are the same regardless of the candidates' performance. In this case, we can think of evaluators behaving as if the revelation of information, through the noisy signal, does not affect their decision. Refusing to use the information about the performance of candidates in group j reflects that evaluators from class $\mathcal{E}$ treat signals as uninformative, but this reaction could be the same for all the groups they are evaluating. Therefore, the key point is to find evidence that evaluators from class $\mathcal{E}$ are not willing to update information for one group but they are willing to do so for another. It is then straight forward to show how to test for willingness to update beliefs as the alternative hypothesis to Test 2.

Test 3 (Willingness to change) *If $q_j^H \neq q_j^L$, evaluators are willing to change their initial beliefs about candidates from group j.*

Under Test 3 evaluators modify their beliefs, hence if we provide them with enough information about the candidates' performance their priors will change. Evaluators for whom $q_j^H \neq q_j^L$ behave as Bayesian players, updating the priors in the presence of information. The data to be used to test this theory in post-apartheid South Africa is presented below.

4 Data sources

The model presented above is simple but general enough so it can be applied in scenarios where agents receive a noisy signal and have a chance to update their priors and then reveal their posterior beliefs. Unfortunately, finding such a dataset presents a challenge. However, I will argue that it is possible to use data from the South African version of the TV show *The Weakest Link* to understand the causes of discrimination: priors and willingness to change them.

4.1 The experiment

The Weakest Link is a winner-takes-all television game where nine participants answer several trivia questions. These participants have a decreasing amount of time to answer as many questions as possible in each round. At the end of a round, each player decides individually, secretly and simultaneously who to vote off the game. When the votes are revealed, the person with the highest number of votes leaves the game. The remaining participants move on to the next round and keep answering questions and eliminating one player per round until two players are left. The player who answers the most questions correctly in the final stage wins. The prize is a function of the number of correct questions throughout the game[10]. Players can win a maximum of R60,000, approximately US$10,000 or US$ 21,200 using PPP.

The game has all the components needed to estimate the model of evaluators and candidates discussed in section 3.1. First, the participants do not know each other before playing the game, which increases the propensity for players to have priors based on observables. As shown in Table 1, most players come from the Johannesburg-Pretoria area, where the show is produced. On the day of the filming, these participants are asked to go the production company. These people do not know each other before that day. Players from other cities

[10]The prize is the amount of "banked" money, and banking is allowed after a correct answer.

are flown in and stay in different hotels, and they also do not know each other. All the participants finally meet when they board the bus that will take them to the studio (a 15 minute ride.) Most of them do not talk to each other during the ride.[11]

Second, players have to identify their opponents' ability to find out who is the weakest link (i.e., the player with the lowest ability), but this is not directly observed. In other words, when choosing who to eliminate players face the same problem as the evaluator in the previous model, while they assume the role of candidate when answering questions. All players see is the other player's observable characteristics such as race, gender and age. Third, "ability" is observed as a noisy signal in the form of the number of questions answered correctly. Answering a question is considered a noisy signal because the questions' difficulty does not vary with the group or each player's performance within a round of the game.[12] The observed performance of each player becomes a random variable. After each answer is provided the show's host indicates whether the answer was correct or not. There is no room for people to interpret the results in different ways. The show's host is the "judge" that defines the quality of the signal. Fourth, at the end of each round, players reveal their posterior probabilities through their voting patterns, which in principle we can assume reflects their choice regarding who they think is the weakest link.

Another advantage of using this game is the prizes, which are much higher than the ones used in experiments. One possible disadvantage is the fact that the sample is not a random draw from the population of South Africa.[13] As shown in Table 1, the demographics do not necessarily match the population distribution.[14]

The Apartheid regime that ended in 1994 with the first multiracial elections created

[11]Personal interview with Duncan Irvine and Kee-Leen Irvine, from Rapid Blue producers of *The Weeakest Link* in South Africa. July 8, 2005.

[12]See footnote 10.

[13]To be in the show players need to first apply. The application is mostly done online, reducing the chances of people from rural areas to be part of the game. Second, the producers at Rapid Blue select the candidates and those selected are asked to take a test. The test is one of general knowledge and according to the producers, these questions have a higher difficulty compared to the ones in the show. Those who pass the test are taken to the studio to see how comfortable they react in front of a camera. Those performing badly are asked to leave. The remaining persons appear in the final broadcast of the show. On average, two shows are taped in a day.

[14]Table 1 shows that the majority of players are white. The producers explained that since the show is broadcast on SABC3, the channel watched by people with higher income, the choice of participants is based on the demographics of the viewers. Also, since the application is mostly done online, there is a high-income bias.

significant differences between races, especially in the accumulation of human capital (e.g Lam and Leibbrandt 2004, Carter and May 2001). To look just at race when there are notable differences in education levels across races would weaken the results, because other variables can be correlated with race. However, because the sample is not random and with a clear urban bias, blacks and whites look very similar on the observables as depicted in Table 2. I will come back to this issue later.

As mentioned in the introduction, this paper differs from Levitt (2004) and Antonovics, Arcidiacono, and Walsh (2005) because to test the implications of this paper I avoid the dynamics of the game, focusing only on the first round where there is no history.[15] We all agree that players would find it optimal to eliminate the weakest players in the early rounds because the prize increases with the number of correct answers.

4.2 The sample

The data is collected from videotapes of three seasons of the show. I prepared a questionnaire to capture the data (available upon request). There are 16-18 shows per season, once we exclude the shows where celebrities play for charity. With three seasons, the sample size has 351 players[16].

The identification of races was done together with a group of South African enumerators. They were asked to indicate whether a contestant they saw on the show was white, black African, coloured (mixed-race), Indian or other. In South Africa, non-white people, including Indians, are included under the word "blacks." For the very few cases where the enumerators disagreed (less than 4%) we played tapes until a consensus was formed. A player was considered "Afrikaner" if he or she was white and the accent sounded like afrikaans. The rest of the players' characteristics were taken directly from the show. Before the host describes the rules of the game, players introduce themselves by saying their name, age, city where they live and occupation. Similar to Levitt (2004) I transform the occupation into an indicator of

[15]For example, vengeance can be a motive for a player to vote off an opponent who voted against her in previous rounds. Also, from round two onwards the player with the highest number of correct questions in the previous round starts the next round, so it is made public who the strongest player is after each round. Finally, once the votes are made public (and before asking the voted off person to leave the game) the show's host interviews two or three participants (at her discretion) asking them about their reasons for their vote changing, which might in turn change the information set of the remainder participants. None of this occurs until after round one.

[16]Some episodes, especially in the third season, are not included due to broadcasting problems

education by inferring the highest level of education needed to perform that job. This was also done with a South African enumerator. These occupations were classified as needing: high school, 2-4 years of college, professional degrees (including a Ph.D.), self-employment, still studying student (college) and unknown (includes housewives, unemployed, retired with unknown previous occupation, and unknown occupation.) Table 1 presents a summary of the statistics.

Table 1: Basic Statistics

Variables	Type	Mean	Median	Std. Dev.
White	binary	0.627	1.00	0.484
African	binary	0.128	0.00	0.335
Coloured	binary	0.105	0.00	0.308
Indian	binary	0.128	0.00	0.335
Afrikaner	binary	0.504	1.00	0.501
Male	binary	33.7	30.0	10.9
Age	years	0.516	1.00	0.500
Johannesburg	binary	0.108	0.00	0.311
Durban	binary	0.188	0.00	0.391
Cape Town	binary	0.208	0.00	0.406
High School	binary	0.359	0.00	0.480
College	binary	0.382	0.00	0.487
Professional degree	binary	0.114	0.00	0.318
Still student	binary	0.066	0.00	0.248
Self-employed	binary	0.017	0.00	0.130
Questions	number	2.75	3.00	0.49
Correct answers	number	2.01	2.00	0.80
Correct answers	proportion	0.734	0.67	0.266
Received a vote	binary	0.387	0.00	0.488
Votes against	number	1.00	0.00	1.75
Sample size: 351 observations				

Whites are overrepresented in the sample, as they account for more than 60% of the participants. As explained above, this is due to the requirements to be on the show and the demographics of the viewers. Two-thirds of the white players were identified as Afrikaners.

Black Africans, coloured and Indians each represent around 12% of the sample. The sample is almost evenly distributed in terms of gender. The players' ages range from 19 to 74, but the sample is biased toward young players, where the median is 30 years of age. Half of the players come from the Johannesburg-Pretoria area where the show is produced, but 10% are from Durban (on the east coast) and 19% from Cape Town (south west.)

The modal player has a job that requires a college degree (38%) and 36% of the sample have jobs that require high school only, with 11% having a job needing a professional-degree. Thus, there is a sample bias towards above-average educated people by South African standards.

Players have two minutes and 50 seconds to answer as many questions as they can in the first round, so the total number of questions varies by participants[17]. The first question is answered by the player with the name's initial is closest to the beginning of the alphabet. The second question is answered by the player to the right of the first respondent (from the TV watchers' view) and so on. The median player answers three questions and very few answer two or four questions. On average players answer two questions correctly and the proportion of correct answers is just above 73%. The second to last row in Table 1 shows that almost 40% of players received at least one vote against them, which indicates that the voting decision exhibits some variability. Finally, the distribution of the number of votes received is skewed to the left, suggesting that votes are not completely diverse.

4.3 Sample selection

The sample of the participants in *The Weakest Link*, as described above, is not a random sample of the population in South Africa. However, while that is a drawback for the extrapolation of the results, it is an advantage for the strength of the estimates. In Table 2, I show that in terms of the observable variables, blacks and whites are similar. The table presents the results from a probit model where the left-hand side variable is the player's race and how this correlates with the observable (demographic) characteristics collected by watching the shows.

An important finding is that whites and blacks do not differ in their educational attainment. If this were not the case differences in race could be due to differences in education.

[17]Hence S is not fixed and for the estimation we use S_k in equations (1) and (2).

Table 2: Probit estimates: Race and observable characteristics

Dependent variable: 1=White 0=Black			
Characteristics	Marginal effect	Std. Dev.	P-value
Male	-0.122	0.054	0.024
Age	0.016	0.003	0.000
Johannesburg	0.083	0.053	0.122
Professional§	-0.048	0.106	0.645
College	0.115	0.077	0.141
High School	0.081	0.078	0.309
Observations: 351		Pseudo R^2: 0.113	
§Beyond college education			

While it is true that being a nuclear scientist might have a different impact on the "evaluators" than a person who is a shopkeeper, on average, these differences do not coincide with race.

It is important to note from Table 2 that whites and blacks differ in their gender composition. However, whites, and not blacks, are the group with a higher proportion of women. Therefore, while gender and race are not independent in the sample, finding evidence of an unfavorable treatment towards black players cannot be explained with an unfavorable treatment towards women. In the next section we explain how to use this data to evaluate the model presented in section three.

5 Estimation

The estimation of the structural parameters of the model $(\alpha_j^0, q_j^L, q_j^H;\ \text{for all } j)$ is done by maximum likelihood. Let $d_{ijk} = 1$ if individual i votes against player k and $d_{ijk} = 0$ otherwise. The likelihood function is then given by

$$\mathcal{L}(\boldsymbol{\alpha^0}, \boldsymbol{q_0}, \boldsymbol{q_1}; \boldsymbol{d}) = \prod_{i=1}^{N_{\mathcal{E}}} \prod_{\substack{k \neq i \\ \forall j}}^{8} P_{ijk}^{d_{ijk}} \tag{4}$$

where P_{ijk} is the probability that player i votes against player $k \in j$ and $\sum_j \sum_k P_{ijk} = 1$. $N_{\mathcal{E}}$ is total number of players in class $\mathcal{E}$ (notice that parameters in **bold** reflect vectors.) Equation (4) is the usual likelihood function when an individual faces multiple (eight) discrete choices, indexed by $k \in j, \forall k \forall j$.

The probability P_{ijk} used here is less common and is computed by appealing to the assumption that, in the first round, players find it optimal to eliminate the participant they believe is the weakest link. In terms of the model described in section 3.1, the probability that player i thinks k is a low-ability player (after seeing k's performance) is given by the posterior α^1_{ijk}. Hence, voting against k can be seen as believing that k is a low-ability player and all other players are not. This is given by

$$p(k \in j \text{ is the low ability player and others are not}) = \alpha^1_{ijk} \prod_{\substack{m \neq \{k,i\} \\ \forall t}} (1 - \alpha^1_{itm}) \tag{5}$$

Because players can vote against one person only (but not themselves), to compute P_{ijk} we need to restrict the probability space to be consistent with this feature of the game. This yields the following expression for P_{ijk}:

$$P_{ijk} = \frac{\alpha^1_{ijk} \prod_{m \neq \{k,i\}, \forall t} (1 - \alpha^1_{itm})}{\sum_s \alpha^1_{its} \prod_{\substack{m \neq \{s,i\} \\ \forall t}} (1 - \alpha^1_{itm})} \tag{6}$$

The likelihood function is obtained by including (6) in (4) with α^1_{ijk} as described in (3). This function is highly nonlinear due to the binomial distribution of Y_{jk} together with the Bayes rule to update the posterior probability in equation (3).

In the multinomial logit framework, it is not possible to identify the parameters (say, β) for each choice. The solution is to restrict one set of parameters to zero. The proof of this result is obtained by adding a non-zero vector (say, λ) to the set of parameters and noticing that $P_{ijk}(\beta + \lambda) = P_{ijk}(\beta)$. It can be shown that this feature is not present in the specification for P_{ijk} described in equation (6) due to the nonlinear relation of the parameter α^0_j, q^H_j, q^L_j and the function α^1_j. Hence, the set of parameters for each choice can be identified.

If we observe only one evaluator assessing only one candidate it will not be possible to disentangle the posterior into priors (α^0_j) and the updating parameters (q^H_j and q^L_j). Identification is achieve via restrictions. First, I assume that evaluator $i \in \mathcal{E}$ treats all candidates

$k \in j$ in the same way, for each j. Second, I assume that all evaluators in class $\mathcal{E}$ behave in the same way. This assumption rules out heterogeneity within member of class $\mathcal{E}$ but provides the variation needed to identify the parameters within j and class $\mathcal{E}$. This creates a trade-off between the gains from identification and the precision of the estimates. Nonetheless, by defining $\mathcal{E}$ ={all players, Afrikaners, other-whites, Africans-Coloured, Indians} I expect the homogeneity assumption to be less severe.

It is also important to mention that I do not have yet an analytical expression to show that the likelihood function defined above is concave over the relevant range of parameters. However, I tried different starting values and the results remained the same. I also verified that in all the estimations the Hessian was a positive semi-definite matrix.

The structural parameters α_j, q_j^H, q_j^L, for j={blacks, whites}, are probabilities and hence limited to take values between zero and one. To guarantee that I assume that each parameter can be expressed as a logistic transformation of a set of *raw* parameters $\psi_j \in \mathbb{R}^3$, as follows[18]

$$\alpha_j^0 = \Lambda(\psi_{1j}) \qquad q_j^H = \Lambda(\psi_{2j}) \qquad q_j^L = \Lambda(\psi_{3j}), \qquad \text{where } \Lambda(a) = \frac{e^a}{1+e^a} \tag{7}$$

The tests derived in section 3.2 will be implemented using the Wald test over the raw parameters described in equation (7). The null hypothesis for Test 1 is that the prior beliefs do not differ by race:

$$\begin{aligned} H_0: &\quad \alpha_{blacks} = \alpha_{whites} \\ H_1: &\quad \alpha_{blacks} \neq \alpha_{whites} \end{aligned} \tag{8}$$

To test for the willingness to update priors I proceed as follows:

$$\begin{aligned} H_0: &\quad q_j^L = q_j^H \\ H_1: &\quad q_j^L \neq q_j^H \end{aligned} \tag{9}$$

Rejecting the null hypothesis implies that priors and posterior will differ, otherwise, evaluators leave their priors unchanged after observing the candidates' signals. The results of these tests will be presented in section 7. But first, I examine whether a player's performance

[18]No restrictions were imposed on the Hessian matrix to produce standard errors within the unit interval.

in the game affects the number of votes he or she receives.

6 Performance and voting patterns

6.1 Performance

Player's performance is heterogenous. Figure 1 shows how the proportion of correct answers varies by race. White players have a higher probability of answering more questions correctly. An average white player answers his or her questions correctly 77% of the time, compared to African (67%), coloured (68%) and Indian (65%) players. While the median white player gets all his/her questions correct, a median black player will answer 2/3 of the questions correctly. Hence, there are clear differences in the performance of players across races. Within the group of white players, the Afrikaner group performs better than the other white players. The median Afrikaner answers all questions correctly while non-Afrikaner white players answer correctly it only 66% of the times, similar to the black players.

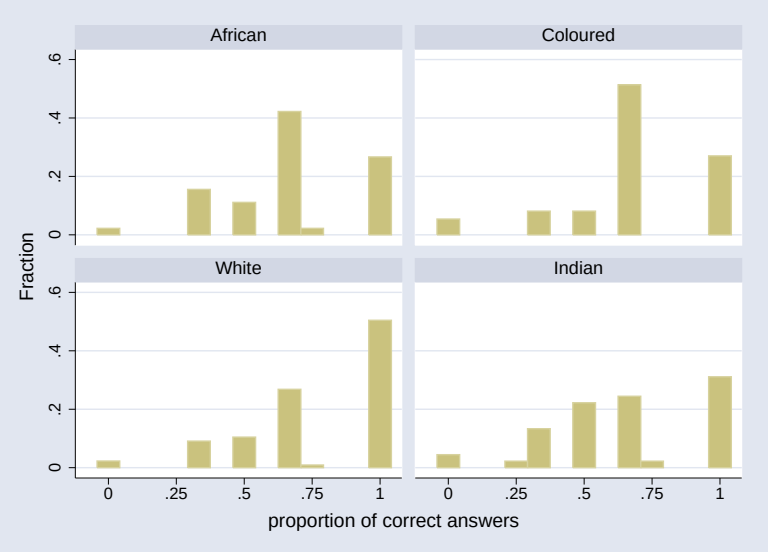

Figure 1: Performance by race

In Table 3 I explore whether this difference in performance by race remains after controlling for other characteristics, including education and gender. I use two different definitions of performance. Column (1) relates players' characteristics with their performance measured

by whether or not the player was the weakest link (i.e. the player with the lowest number of correct answers). None of the players' characteristics correlates with such a measure of performance, except for being Indian. Indian players have a higher probability of having the lowest number of correct answers of the game, after controlling for age, gender, education, etc.

In column (2) I measure performance as the proportion of correct answers used in Figure 1 above. Because the left-hand side variable takes values between zero and one, I use a tobit model to account for an upper and lower limit. As before, the estimates in Table 3 report the marginal effects. I find that whites perform better even after controlling for education and gender. As before, I also find that blacks have a lower performance compared to other races. However, the difference between Afrikaner and other white players no longer exists.

6.2 Reduced form approach

I now turn to issue of voting. In Table 4 I analyze two indicators to measure the voting behavior. Columns (1) to (3) show how the probability of receiving at least one vote depends on the characteristics of the player that is voted against using a probit model (in terms of marginal effects). In column (1) it is observed that black African or coloured players tend to have a higher probability of receiving votes against them. The same is true for younger players. But from the previous results we know that players' performance is associated with race. In columns (2) and (3) I repeat the estimation controlling for whether the player was the worst contestant (column 2) and by the proportion of questions answered correctly (column 3.) The results show that the players' performance, regardless of how is measured, is taken into account when deciding whom to vote against. However, that is not the only information taken into account. Black African or coloured players have a higher probability of being voted against even after controlling for their performance of the game. This holds for the case of younger players too, but not regarding the player's gender.

Columns (3) to (6) repeat the experiment using the number of votes a person receives. In this case the left-hand side is a "count" variable so the estimation is done with a Poisson distribution. The results obtained are analogous to the ones found using the probit model. Again, the players' performance is a good predictor of the number of votes received. However, performance is not the only predictor. The player's race is also important, meaning that

Table 3: Performance and players characteristics (Marginal effects)

Model :	(1)	(2)
Male	-.054	-.003
	(.041)	(.017)
African or Coloured	.084	-.043
	(.061)	(.022)
Indian	.200	-.059
	(.092)	(.031)
Afrikaner	.028	.007
	(.060)	(.022)
Johannesburg	-.018	.026
	(.058)	(.022)
Durban	-.044	.024
	(.071)	(.033)
Cape Town	.045	.009
	(.073)	(.027)
College	-.022	.007
	(.046)	(.019)
Professional degree	.003	-.002
	(.068)	(.028)
Still studying	-.027	-.002
	(.076)	(.038)
Self-employed	.192	-.020
	(.205)	(.066)
Age	-.013	.007
	(.012)	(.005)
Age-squared	.013	-.007
	(.014)	(.007)
Nobs.	351	351
(Pseudo) R^2	0.042	0.045

(1) Probit: 1=Weakest player.
(2) Tobit: Proportion of correct answers.
Note: Standard deviations in parenthesis.

black players receive more votes than white or Indian players with the same performance, regardless of how performance is measured.

Two conclusions can be derived from this section: (1) Players use the information revealed during the game when they make their voting decision and (2) that information is not sufficient because the player's race is also relevant. These facts reveal evidence of discrimination but it is not clear how these two facts can explain what type of discrimination is taking place. The reader might have noticed that these facts are equivalent to "Lakisha's problem" described in the introduction: resumes from black candidates had a higher probability of being rejected even after controlling for credentials (Bertrand and Mullainathan 2004).

Two conflicting hypothesis can explain this facts. *Hypothesis 1*: since players' performance "predicts" the voting behavior it could indicate that players update their priors; thus, it is the differences in priors that explains why race matters after controlling for performance. *Hypothesis 2*: players might be willing to change their beliefs for whites, making performance a good predictor of voting, and not willing to change beliefs about blacks. Because we cannot separate these two potential explanations, having a behavioral model such as the one presented in section 3.1 is important to identify these two hypotheses. In the next section I present the estimates of the proposed model.

7 Priors and willingness to update

Recall that the goal of the paper is to first estimate evaluator's priors (α_j^0) about the candidate's race. The second goal is to test whether or not evaluators are willing to update their priors after observing noisy signals from candidates. When priors differ by the candidates' race it is considered evidence of negative stereotypes. When evaluators update priors in different ways for different races they are also discriminating. Below I present the results of the estimation followed by simulations and robustness checks.

7.1 Estimates

I first evaluate the priors and test for difference across races. Table 5 shows the estimates for the parameter α_j^0 when $j = \{\text{blacks,whites}\}$.

The first column presents the estimates using the full sample. The point estimates show

Table 4: Votes received and players characteristics

Model :	(1)	(2)	(3)	(4)	(5)	(6)
Intercept				2.311	1.315	3.770
				(0.653)	(0.679)	(0.664)
Male	-.023	.005	-.023	-0.058	0.100	-0.142
	(.055)	(.058)	(.061)	(0.112)	(0.115)	(0.114)
African or Coloured	.200	.180	.182	0.486	0.281	0.425
	(.070)	(.075)	(.079)	(0.133)	(0.133)	(0.128)
Indian	.137	.054	.036	0.435	0.057	0.208
	(.097)	(.101)	(.108)	(0.185)	(0.198)	(0.185)
Afrikaner	-.016	-.029	-.004	0.044	-0.017	0.030
	(.072)	(.076)	(.082)	(0.159)	(0.160)	(0.161)
Johannesburg	-.049	-.050	-.014	-0.190	-0.159	-0.026
	(.073)	(.077)	(.080)	(0.150)	(0.151)	(0.152)
Durban	-.066	-.053	-.035	-0.283	-0.199	-0.220
	(.104)	(.112)	(.120)	(0.235)	(0.245)	(0.236)
Cape Town	-.062	-.093	-.070	0.164	0.033	0.036
	(.085)	(.089)	(.094)	(0.168)	(0.172)	(0.174)
College	.045	.063	.079	-0.259	-0.180	-0.326
	(.061)	(.064)	(.069)	(0.129)	(0.131)	(0.132)
Professional degree	.086	.104	.104	0.151	0.107	0.172
	(.095)	(.102)	(.108)	(0.166)	(0.169)	(0.167)
Still studying	-.155	-.167	-.180	-0.363	-0.345	-0.519
	(.103)	(.111)	(.104)	(0.217)	(0.220)	(0.227)
Self-employed	.018	-.115	-.040	-0.376	-0.741	-0.198
	(.211)	(.211)	(.220)	(0.512)	(0.513)	(0.512)
Age	-.043	-.039	-.036	-0.118	-0.101	-0.099
	(.017)	(.018)	(.018)	(0.033)	(0.035)	(0.034)
Age-squared	.049	.045	.045	0.130	0.119	0.129
	(.021)	(.022)	(.022)	(0.039)	(0.043)	(0.042)
Weakest link		.532			1.740	
		(.057)			(0.112)	
Prop. questions correct			-1.312			-3.347
			(.136)			(0.197)
Nobs.	351	351	351	351	351	351
(Pseudo) R^2	0.04	0.161	0.300	0.045	0.244	0.290

1)-(3) Probit: 1=Player received a vote against, marginal effects
(4)-(6) Poisson: Number of votes received
Note: Standard deviations in parenthesis.

that the average player associates a prior probability that a black contestant is a low-ability type to be equal to 72%, while the same prior probability for whites is 74%. There are interesting differences on how these perceptions change by groups of evaluators. Afrikaners and African players have a prior that favors whites. The opposite is found for non-Afrikaner whites and Indians.

Table 5: Estimates for priors (α_j^0), by groups

	Evaluators				
Parameters	All	Afrikaner	Other Whites	Africans & Coloured	Indian
α^0_{blacks}	.721	.827	.676	.772	.362
	(1.58)	(.499)	(1.15)	(.660)	(.632)
α^0_{whites}	.741	.703	.776	.712	.861
	(1.51)	(.728)	(.912)	(.769)	(.327)
Wald test	.016	.743	.352	.146	8.59
p-value	.900	.389	.553	.703	.003
Nobs.	351	70	150	82	45

Note: Standard deviations in parentheses. Max. likelihood estimates
$H_0 : \alpha_j^0 = \alpha_t^0$. Critical value: $\chi^2_{95\%}(1) = 3.84$

The bottom panel of the table shows the test when the null hypothesis is of equal priors. I cannot reject the null hypothesis, suggesting that there are not statistical differences in priors across races. These results hold for all groups with the exception of Indians. For the case of the Indians we have to consider the differences in sample size. Indians alone account for 13% of the sample. When I considered the priors from all blacks (not included in table 5), I still cannot reject the null hypothesis of equal priors.

I now turn to the other two structural parameters: q_j^H and q_j^L. In Table 6, I show these estimates for the whole sample and then for different subgroups of evaluators. Recall that q_j^H measures the likelihood that a good signal comes from a high ability player, while q_j^L refers to the corresponding likelihood as coming from a low ability player. For the full sample and across groups I found that $q_j^H > q_j^L$ for $j =$\{blacks, whites\}, suggesting –as expected- that a

high ability player will have a higher chance of answering a question correctly. For example, in the full sample, q^H_{black}=.949 and q^L_{black}=.789.

Table 6: Willingness to update priors, by groups

	Evaluators				
Parameters	All	Afrikaner	Other Whites	Africans & Coloured	Indian
q^H_{blacks}	.949	.905	.985	.914	1.00
	(.052)	(.181)	(.064)	(.154)	(.010)
q^L_{blacks}	.789	.701	.933	.681	.997
	(.186)	(.461)	(.277)	(.442)	(.069)
Wald test	2.22	2.00	1.32	3.38	2.81
p-value	.136	.157	.251	.066	.094
q^H_{whites}	.912	.934	.898	.924	.774
	(.073)	(.128)	(.141)	(.120)	(.242)
q^L_{whites}	.659	.704	.661	.672	.360
	(.214)	(.452)	(.361)	(.397)	(.340)
Wald test	2.71	3.00	2.06	3.20	4.09
p-value	.100	.083	.151	.074	.043
Nobs.	351	70	150	82	45

Note: Standard deviations in parentheses. Max. likelihood estimates
$H_0 : q^H_j = q^L_j$. Critical value: $\chi^2_{95\%}(1) = 3.84$

The table also shows the tests for willingness to change (or update) prior beliefs. Recall that when $q^L_j = q^H_j$ the prior and the posterior will be identical, meaning that information is not relevant. On the other hand, rejecting the null hypothesis of $q^L_j = q^H_j$ is taken as evidence of a behavior where priors are updated. For all players, I cannot reject the null hypothesis of not-updating regarding black participants. The null hypothesis is on the margin regarding white players.

The interesting results appear when we look across groups. White players (Afrikaners and other whites) are not willing to change their beliefs about black players. I cannot reject the null hypothesis that $q^L_j = q^H_j$. They seem to behave as if information is not important

when evaluating a black contestant. But Afrikaners differ from other white players by being willing to update beliefs regarding other whites. Non-Afrikaner whites are not willing to do so for any race. On the other hand, blacks behave differently. Indians, Africans and coloured players are willing to change beliefs about other blacks, but they are also willing to change for white players. These results suggest that players do take into account other participants' performance but it depends on which player they are evaluating and who is making the evaluation. For blacks the results suggest that their voting pattern takes into account the performance of all players. For non-Afrikaner whites, the participants' performance is not relevant at all and for Afrikaners, performance is valid when evaluating white players but not black ones. Hence, difference groups seem to have different behaviors depending on who they are evaluating.

7.2 Discussion

I considerer the fictitious case where four evaluators: two whites –an Afrikaner and a non-Afrikaner–, a black African and an Indian; have to choose between two candidates. One candidate is black and the other one is white. The goal is to use the estimates for α_j^0, q_j^H and q_j^L from tables 5 and 6 to simulate the evaluators' posterior probability that the candidates are low-ability types for different number of correct answers (Y_{jk}) that candidates can have. The total number of questions asked (S) is set to be equal to three for all candidates.

Because I do not compute the standard errors for the simulated posteriors, the estimates shown in Table 7 take the point estimates from the previous tables only for the cases where the null hypothesis (of equal prior or equal q's) is rejected. For all other cases, when two parameters are statistically equal to each other they are replaced by their average. For example, I showed above that we cannot reject the null hypothesis that Afrikaners have the same prior for blacks and whites (Table 5, second column). The point estimate for blacks is .827 and for whites is .703. For the simulations I use .765 for both parameters. Since this group of evaluators also does not update their priors for blacks (I cannot reject the null hypothesis that $q^H_{blacks} = q^L_{blacks}$) then we use $q^H_{blacks} = q^L_{blacks} = .959$, the average of .905 and .701 (Table 6, second column).

When evaluators are willing to change prior beliefs, the simulated posteriors decrease as the number of corrects questions increases since the posterior is about the probability that

Table 7: Simulated posterior (α_j^1). By groups

	Evaluators							
	Afrikaner		Other Whites		Africans		Indians	
Candidates:	black	white	black	white	black	white	black	white
Prior (α_j^0)	0.765§	0.765§	0.726§	0.726§	0.742§	0.742§	0.362	0.861
q_j^H	0.803†	0.934	0.959†	0.780†	0.914	0.924	1.000	0.774
q_j^L	0.803†	0.704	0.959†	0.780†	0.681	0.672	0.997	0.360
Correct answers	Posterior probabilities							
$Y_{jk} = 0$	0.765	0.997	0.726	0.726	0.993	0.996	1.000	0.993
$Y_{jk} = 1$	0.765	0.980	0.726	0.726	0.967	0.975	1.000	0.959
$Y_{jk} = 2$	0.765	0.892	0.726	0.726	0.856	0.868	1.000	0.791
$Y_{jk} = 3$	0.765	0.582	0.726	0.726	0.543	0.525	0.360	0.384
Votes against	black		any		black		white	

Estimates based on parameters from tables 5 and 6. Total number of questions $S = 3$
§†Average of estimated parameters

the k-th contestant is a low-ability type. Consider the case when both candidates respond to all three questions correctly ($Y_{jk} = 3$). The Afrikaner evaluator would tend to vote against the black candidate because the posterior for blacks is higher (.765) than the one for whites (.582) even when the prior was the same for both. Such a black candidate will also be voted against by the African evaluator and there is a 50% chance that the other white evaluator would vote also against the black candidate. Only the Indian evaluator would vote against the white candidate. In half of the cases the black candidate will get three votes against him out of possible four. Black candidates with good signals will be eliminated instead of equally-performing white candidates. These results are driven by the unwillingness to update beliefs from the part of the evaluators.

7.3 Robustness of the estimates

To confirm the robustness of the above estimates (and the simulations), I redo the estimation excluding from the sample the players with the worst performance. The idea is that the set

of "candidates" that players with the lowest performance face are different compared to the other players. This could be considered as a way to explore the role of heterogeneity within a class of evaluators.

In Table 8 I use two different measures of low performance: lowest proportion of correct answers and lowest number of correct answers. The results show that blacks still update beliefs for all players and whites continue refusing to do so for blacks. The test for white players updating for other white players is on the margin at the 10% significance level for the null hypothesis of no updating. These results confirm the robustness of the estimates using all the players in the sample.

Table 8: Testing willingness to update without worst players

Race	All	Afrikaner	Other Whites	Africans & Coloured	Indian
Lowest proportion of correct answers					
Blacks (p-value)	.218	.284	.313	.010	.076
Whites (p-value)	.172	.116	.263	.070	.095
Nobs.	253	60	114	53	25
Lowest number of correct questions					
Blacks (p-value)	.202	.262	.316	.013	.087
Whites (p-value)	.126	.120	.168	.024	.211
Nobs.	288	62	130	63	32
$H_0 : q_j^H = q_j^L$. Crit. value: $\chi^2_{95\%}(1) = 3.84$					

8 Conclusions

This paper introduces a model of evaluators and candidates where discrimination can occur due to two reasons. First, evaluators can have negative stereotypes against a group of candidates. Second, after observing signals from candidates, evaluators might decide to use those signals differently for different groups of candidates. This differential treatment is a second source of discrimination, a refusal to let relevant information disturb prior beliefs.

One contribution of this paper is to provide a unified approach for two sets of models that were, until now, providing partial explanations for the observed discriminatory behavior.

Having such a model is crucial because it allows us to go beyond finding evidence of discrimination and more into its sources. This, in turn, permits a better development of anti-discrimination policies. Another important contribution of the paper is the development of testable implications that allow us to contrast the model using data.

By using data from the South African version of the television show *The Weakest Link* the paper finds evidence of discrimination against black "candidates." The source of discrimination is not the existence of negative priors against blacks but the fact that white players behave as if they refuse to use information in order to asses the quality of black candidates. Whites may not have different priors for blacks and whites, but for blacks they are not willing to change them. This behavior is the source of discrimination.

From a theoretical point, the paper models only the behavior of evaluators. A natural extension for the model is to include how the candidates' decisions on human capital are affected by the behavior of the evaluators. This is left for future research.

The use of a data from a TV show limits how generalizable the results of the study are. In the absence of an experiment drawn from a more representative sample of the population, the current results shed some light about the process undertaken by individuals when they have incomplete information about other people's ability. Nonetheless, a third contribution of the paper is to show that discrimination can occur in the absence of overtly negative priors. Using observable characteristics to infer unobservable ones leads to an unequal treatment of individuals. Social psychology suggests that the use of stereotypes is an inevitable process. It is what we do when we do not know. However, refusing to use information on individuals from a group, but not for another, is a deeper form of discrimination. Not to wish to know may indeed be worse. Finding policies to overturn such behavior is a pending issue.

References

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

Altonji, J., and R. Blank (1999): "Race and Gender in the Labor Market," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 3, pp. 3144–3259, Amsterdam. North-Holland.

Anderson, L. R., R. G. Fryer, and C. A. Holt (2005): "Discrimination: Experimental Evidence from Psychology and Economics," in *Forthcoming Handbook on Economics of Discrimination*, ed. by W. Rogers.

Antonovics, K., P. Arcidiacono, and R. Walsh (2005): "Games and Discrimination: Lessons From The Weakest Link," *Journal of Human Resources*, forthcoming(Fall).

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

Ayres, I., and P. Siegelman (1995): "Race and Gender Discrimination in Bargaining for a New Car," *American Economic Review*, 85(3), 304–21.

Banaji, M. R. (2002): "Stereotypes, social psychology of," in *International Encyclopedia of the Social and Behavioral Sciences*, ed. by N. Smelser, and P. Baltes, pp. 15100–15104, New York. Pergamon.

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

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.

Carter, M. R., and J. May (2001): "One Kind of Freedom: Poverty Dynamics in Post-apartheid South Africa," *World Development*, 29(12), 1987–2006.

Casale, D. (2003): "The Rise in Female Labour Force Participation in South Africa: An Analysis of Household Survey Data, 1995-2001," Department of economics, University of Natal.

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

Darity, William A, J., and P. L. Mason (1998): "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender," *Journal of Economic Perspectives*, 12(2), 63–90.

Fiske, S. T. (1998): "Stereotyping, prejudice, and discrimination," in *The Handbook of Social Psychology*, ed. by D. T. Gilbert, S. T. Fiske, and G. Lindzey, pp. 357–411, New York. McGraw Hill.

Frijters, P. (1999): "Hiring on the Basis of Expected Productivity in a South African Clothing Firm," *Oxford Economic Papers*, 51(2), 345–54.

Goldin, C., and C. Rouse (2000): "Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians," *American Economic Review*, 90(4), 715–741.

Lam, D., and M. Leibbrandt (2004): "What's happened to inequality in South Africa since the end of apartheid?," Manuscript, University of Cape Town.

Lang, K. (1986): "A Language Theory of Discrimination," *The Quarterly Journal of Economics*, 101(2), 363–82.

Levitt, S. D. (2004): "Testing Theories of Discrimination: Evidence from Weakest Link," *Journal of Law and Economics*, XLVII(2), 431–52.

Lundberg, S. J. (1991): "The Enforcement of Equal Opportunity Laws under Imperfect Information: Affirmative Action and Alternatives," *The Quarterly Journal of Economics*, 106(1), 309–26.

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

Moreno, M., H. Ñopo, J. Saavedra, and M. Torero (2004): "Gender and Racial Discrimination in Hiring: A Pseudo Audit Study for Three Selected Occupations in Metropolitan Lima," Discussion Paper 979, Institute for the Study of Labor (IZA).

Neumark, D. (1996): "Sex Discrimination in Restaurant Hiring: An Audit Study," *The Quarterly Journal of Economics*, 111(3), 915–41.

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

Psacharopoulos, G., and H. Patrinos (1994): *Indigenous People and Poverty in Latin America: An Empirical Analysis*. The World Bank, Washington, DC.

A Preference versus information-based discrimination

The papers by Antonovics, Arcidiacono, and Walsh (2005) and Levitt (2004) try to distinguish between discrimination based on preferences and discrimination based on information. The first theory comes from Becker (1957). Becker explains discrimination as related to individual's preferences or tastes. These individuals prefer not to interact with those discriminated against." As Becker explains "[i]gnorance may be quickly eliminated by the spread of knowledge, while prejudice (i.e, preference) is relatively independent of knowledge." (p. 16) He continues

> "Many prejudiced people often erroneously answer questions about groups they discriminate against; their 'ignorance' about these groups, however, is of secondary importance for understanding and combating their discrimination, since their behavior is independent of all attempts to give them the facts." (p. 16, n. 4)

The second theory comes from the work of Arrow (1973), Phelps (1972) and is extended by Coate and Loury (1993). In these models employers observe signals from workers and discrimination is explained by negative stereotypes against a group of workers. This approach "can be thought of as reflecting not tastes but perceptions of reality." (Arrow 1973, p. 23.) Here people use group identity, such as race, gender or age, as a proxy for unobserved ability. But when information is provided, their initial belief will change accordingly. Because this approach relies on the information available to employers, it has been labeled "information-based" discrimination.

It is possible to distinguish between these two models in a way that is different from what Antonovics, Arcidiacono, and Walsh (2005) and Levitt (2004) have done. We can do this by testing Becker's statement about how people discriminating based on preferences would react in the presence of information. Providing these individuals with information about the productivity of those suffering from discrimination will not change their discriminatory behavior. They behave as if they are unwilling to change their prior beliefs or negative stereotypes.

In terms of the model introduced in this paper, when evaluators do not change their priors ($q_j^H = q_j^L$) behavior would be consistent with that of prejudiced people described by Becker. Otherwise, when $q_j^H \neq q_j^L$, evaluators are willing to change their priors behaving as the agents in Arrow's (1973) model. However, it is not clear how the findings of this paper –evaluators having the same prior for all candidates– could be understood according to these two models.