View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Universidad Carlos III de Madrid e-Archivo

UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 08-06
Business and Economics Series 1
February 2008

Departamento de Economía de la Empresa
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-96-07

Why do I like people like me?*

Manuel F. Bagues[1] and María José Pérez Villadóniga[2]

Abstract

In many dimensions the ability to assess knowledge depends critically on the observer's own knowledge of that dimension. Building on this feature, this paper offers both theoretical and empirical evidence showing that, in those tasks where multidisciplinary knowledge is required, evaluations exhibit a similar-to-me effect: candidates who excel in the same dimensions as the evaluator tend to be ranked relatively higher. It is also shown that, if races or genders differ in their distribution of ability, group discrimination will arise unless evaluators (i) are well informed about the extent of intergroup differences and (ii) they may condition their assessments on candidates' group belonging.

JEL codes: J71, D82
Keywords: Statistical Discrimination, Evaluation Biases.

* We would like to thank Antonio Cabrales, Juan Jose Dolado, Florentino Felgueroso, Walter García-Fontes, Juan F. Jimeno, Günter Fink, Marco Haas, Franck Malherbet, Pedro Marín, Eduardo Melero, Nicola Pavoni and Enrico Pennings for many useful comments on an earlier version of this paper. We are also grateful to participants in seminars at the universities of Bocconi, Verona, Pavia, Las Palmas de Gran Canaria, Oviedo, León, Sant'Anna School of Advanced Studies, FEDEA and conference presentations at the Jornadas de Economía Industrial, Simposio de Análisis Económico and SOLE-EALE. We also thank the production manager of The Weakest Link for kindly providing information about the functioning of the show. Financial support from the project MEC-04-SEJ2004-04065 is gratefully acknowledged by the second author. Any remaining error is our own.

[1] mfbagues@emp.uc3m.es. Universidad Carlos III, Departamento de Economía de la Empresa, Calle Madrid 126, 28903 Getafe, Madrid, Spain.

[2] mjpvilla@uniovi.es. Universidad de Oviedo, Departamento de Economía, Avenida del Cristo, s/n, 33006 Oviedo, Spain.

1 Introduction

The fact that individuals might be treated differently according to exogenous characteristics such as gender, age or race has been well documented in the literature. Most of the evidence refers to the labour market, where differences in wages or hiring and promotion that cannot be accounted for by differences in productivity have been observed.[1] Discriminatory behaviours have also been observed in housing decisions (Massey and Denton 1993), lending (Hunter and Walker 1996), car selling (Ayres and Siegelman 1995) or even in the refereeing of academic papers (Blank 1991; Fisher et al. 1994).

In the economics literature, two distinct general sets of explanations that focus on the demand side of the labour market have been proposed to explain the origin and persistence of discrimination. On the one hand, taste models, as in Gary Becker's (1957) seminal work, suggest a preference-based motivation for the existence of discrimination. The difference in wages between two equally productive groups of workers arises because employers, customers or co-workers dislike interacting with employees that belong to certain groups. On the other hand, statistical models of discrimination argue that, in the presence of information asymmetries about the real productivity of workers, the group-belonging of an individual can be considered as a signal that provides additional information. Groups of workers may differ in their expected productivity (Phelps 1972, Lazear and Rosen 1990) or in the reliability of the observable signals (Aigner and Cain 1977, Cornell and Welch 1996). In this context, taking into account an individual's group affiliation may be a rational response to its informational content and the wage gap might persist in the long run.

Independently of its origin, in general it has been considered that discriminatory outcomes will decrease if hiring procedures do not allow evaluators to observe or to take into account the identity and group belonging of candidates (as in Blank 1991 or Golding and Rouse 2000). In this paper we show that, in positions that require multidisciplinary knowledge, this is not true any longer. On the contrary, in such framework a discriminatory outcome might arise unless candidates' group belonging can taken into account by evaluators.

In particular, within the framework of statistical discrimination models, we propose a model which builds on the following three assumptions. First, productivity is assumed to be multidimensional. While in previous models of statistical discrimination productivity has been typically considered unidimensional,[2] as Aigner and Cain (1977, page 176) acknowledge evaluating workers' ability "undoubtedly involve(s) a number of measures". In this paper we allow for the existence of multiple dimensions of ability that can be understood either as different tasks that the worker needs to undertake, or as separable skills that are required to perform a single task. Second, while standard models of statistical discrimination assume that the accuracy with which employers assess the productivity of potential employees is exogenous or depends on the group

[1] For a survey see, for instance, Altonji and Blank (1999).

[2] See Phelps 1972, Arrow 1973, Aigner and Cain 1977 or Cornell and Welch 1996.

belonging of the employer and the candidate,[3] here we will assume that the capability of an employer to evaluate an individual's quality at a certain dimension increases with her knowledge of that dimension. This assumption is consistent with experimental evidence, where it has often been found that, in many fields, individuals who are less competent are also less accurate at evaluating ability than competent ones.[4] Finally, the third feature of the model is that, while all groups of candidates are, on average, equally productive, their quality is not necessarily the same in every dimension.

Combining these features our model yields the following two predictions. First, we show that a similar-to-me-in-skills effect arises in the evaluation. Since individuals can assess knowledge more accurately at those dimensions where they are more knowledgeable, an employer who makes an optimal use of the available information will give relatively more weight to those signals that are observed in dimensions where she is most knowledgeable. As a result, given any two equally productive candidates, the employer will tend to give a higher valuation to the candidate who excels in the same dimensions as she does. This result is consistent with the evidence found by Bagues et al. (2007) who observe that, in public exams in Spain, evaluators take most into account candidates' performance in those fields in which they are themselves relatively more knowledgeable. As a consequence, candidates have more chances of success if they are (randomly) assigned to an evaluation committee whose members are specialized in those fields where they excel. More generally, the fact that evaluators tend to give higher ratings to candidates who are similar to themselves has been widely documented both in psychology and sociology (Byrne 1971) and in organizational processes, such as supervisors' assessments of subordinates or recruitment (Goldberg 2005). Note, however, that while in these studies similarity is understood as similar attitudes or observable personal attributes, in this case we argue that similarity in terms of skills may also be relevant.

Second, the model shows that when (equally productive) groups differ in their distribution of ability across dimensions, the existence of a similar-to-me-in-skills effect may generate group discrimination.[5] In particular, group discrimination will arise if (i) employers are not fully aware of the extent of these differences or (ii) employers are perfectly informed but cannot condition their evaluations on candidates' group-belonging. The intuition behind this result is the following. Employers will tend to give more weight to signals that have been observed in those dimensions where they are more knowledgeable. In principle this favours candidates belonging to the same group as the employer,

[3] Ibid.

[4] Knowledgeable people are more accurate in their evaluations in the field of chess (Chi 1978), physics (Chi et al. 1982), grammar (Kruger and Dunning 1999) or academic performance (Everson and Tobias 1998). Similarly, in the context of firms R&D strategies, Cohen and Levinthal (1990) argue that "the ability of a firm to recognize the value of new, external information, is largely a function of the firm's level of prior related knowledge".

[5] Following Aigner and Cain (1977), we consider group discrimination as the situation where "groups that have the same average ability may receive different average pay" (pp.178). Note that in a multidimensional framework the term *same ability* should be interpreted as meaning *same total ability* rather than *same ability at every dimension*.

as they are more likely to excel precisely in these dimensions. Still, a well-informed evaluator who was allowed to take into account the group belonging of candidates might use this information in order to adjust her priors. This would not only be efficient from an informational point of view but, as well, it would yield similar average evaluations across groups of candidates.

There are many cases where such correction might not be feasible. Sometimes employers might not be fully aware of the fact that equally productive groups differ in terms of how good they are in each dimension. This may happen when groups have little interaction, when the size of the minority is relatively small[6] or in the presence of a number of cognitive biases such as observational selection bias, availability bias or anchoring that can generate a divergence between the individual's perception of other groups' quality profiles and their true quality distribution. As well, even if evaluators are well informed, they might not be allowed to take the group belonging of candidates into account. This may be the result of antidiscrimination laws, according to which, sometimes, candidates' identity is kept anonymous or evaluators are explicitly instructed not to take candidates' group belonging into account. Paradoxically, our results suggest that, if groups differ in their distribution of quality across dimensions, candidates that belong to the employers' group will tend to preferred unless group belonging is taken into account.

Our model overcomes some of the drawbacks of standard models of statistical discrimination (Aigner and Cain 1977, Cornell and Welch 1996). First, it relies on more plausible assumptions. Models of statistical discrimination typically assume that the screening technology is exogenous or depends on the group belonging of the candidate. In contrast, in our setup the accuracy of the evaluation depends on the evaluator's knowledge of each dimension, all groups of candidates being evaluated with the same expected accuracy. This is consistent with an abundant literature which finds a positive relationship between evaluators' knowledge of a field and the quality of their assessment.[7] On the contrary, empirical evidence supporting that the accuracy of evaluations varies over groups of candidates is very scarce.[8] Second, note that the predictions of the model also differ. The above models of statistical discrimination predict that, among highly productive candidates, those belonging to the evaluator's group tend to be preferred. However, the reverse does not hold and, when all candidates are relatively unproductive, those who do not belong to the employer's group are favoured, given that the observed (low) signal about their quality is a weaker indicator of their productivity. This is an uncomfortable prediction, as there is no evidence supporting the reversal of the race and gender gap for low productivity levels. In contrast, in this multidimensional framework the evaluation bias favours those candidates akin to the evaluator for every level of productivity.

The empirical testing of the model proposed in this paper presents very de-

[6] As it would increase the cost of rationality. See for instance Fryer and Jackson (2007).

[7] See footnote 4.

[8] Up to our knowledge, and according to Holzer and Neumark (2000), Neumark (1999) provides the only empirical analysis about the reliability of the labor market information of various groups.

manding information requirements. First, it requires data, at each dimension, on the quality of evaluators and candidates. Second, it needs information on the evaluations performed. Such data is rarely observable in the labour market. Moreover, in most observational studies it is not possible to assure that all agents observe the same informational set. To avoid these caveats, we test the validity of the proposed model by exploiting the evidence provided by a well-known TV contest. In particular, we use data from The Weakest Link TV show. In this contest players answer a number of questions over a series of rounds. At the end of each round players are asked to vote who was the weakest link or worst contestant of the round. The player who gets the most votes is eliminated. A convenient feature of the Spanish version of the show is that questions are explicitly classified in different fields. Therefore, we have a context where productivity is multidimensional -players quality depends on how knowledgeable they are across a number of different fields- and where it is possible to observe individuals' performance at each dimension and the assessment received. Furthermore, since all contestants are candidates and evaluators at the same time, we can also observe evaluators' knowledge profile.

As predicted by the model, we observe the existence of a similar-to-me-in-skills effect. Contestants tend to give a higher valuation to those participants' who excel in the same dimensions as themselves. We also observe that quality distribution is correlated within age and gender groups. Interestingly, the evidence shows that players take into account in their evaluations the existence of these differences across groups. Doing so is not only informationally efficient but, moreover, it prevents the existence of a bias in favour of candidates that belong to the same group as the evaluator. Our analysis thus provides unique evidence documenting how, in some cases, allowing evaluators to take into account in their assessments candidates' group belonging may actually prevent a discriminatory outcome.

Our papers adds to a number of studies that have previously analyzed The Weakest Link TV show. Février and Linnemer (2006) study players' optimal voting strategy using French data. Levitt (2004) and Antonovics et al. (2005) use data from the American version of the show to empirically distinguish between different competing discrimination theories. In another paper, Antonovics et al. (2008) use these data to determine whether the performance of contestants is affected by the gender of their opponents. Finally, using the British version, Haan et al. (2004) have pointed out the limitations in the rationality of players' banking decisions.

The paper is structured as follows. In the remaining of this section we offer an example that helps to clarify the intuition underlying our model. Additionally, another example is provided in the appendix. The formal model and the empirical implications are discussed in Section 2. Section 3 describes the structure of The Weakest Link contest and the available data. Section 4 presents the empirical evidence and Section 5 concludes.

1.1 Example: The Academic Job Market

Every year, PhD candidates in Economics attend the academic job market. There, candidates meet with employers who seek young promising researchers that might be willing to join their faculty. A good candidate is understood as someone who will be able to make a relevant scientific contribution, typically in the form of an academic publication. However, the true quality of a candidate cannot be immediately disclosed and several years may be necessary in order to assess the importance of the contribution. In the meantime, the main signal that the evaluator can observe about the candidate's quality will be her job market paper.[9]

Evaluating a job market paper is in general a multidimensional task. First, a good paper should provide an interesting and novel economic idea. Second, it is usually expected that this idea is presented through an elegant mathematical formalization. Evaluating each of these dimensions is likely to be complex and evaluators usually will be able to assess more accurately the quality of the paper in that dimension where they are themselves more knowledgeable[10]. For instance, while evaluating the novelty of a paper requires knowledge of the previous related literature, in order to appreciate its mathematical quality being skilled in mathematics might be highly convenient.

Following with the example, let us assume for simplicity that the total quality of a paper is equal to the sum of its quality in the above two dimensions: its quality in terms of economic novelty and relevance plus its quality at the mathematical dimension. Let us also imagine the extreme case of an evaluator who is perfectly knowledgeable in mathematics but is completely ignorant of the literature. In this case, given the evaluator's incapability to appreciate the economic novelty of the paper, an optimal evaluation would involve taking mostly into account the information that she observes along the mathematical dimension. Therefore, faced with two papers of identical quality, the evaluator would tend to give a higher valuation to the paper which excels relatively more in the mathematical dimension.

Moreover, imagine that there exist gender differences in the distribution of ability across dimensions. Consider, for instance, a scenario where the (total) quality of papers is independent of the author's gender but, still, where male economists tend to have a higher capability for mathematical abstraction, while female economists are better in terms of the economic relevance of their work. Then, the above similar-to-me-in-skills effect could also generate gender discrimination. The intuition is the following. If evaluators are unaware of the existence of these gender differences or, even if they are aware, they are not allowed to condition their assessments on candidates' gender, then the evaluators' opti-

[9] As Levinovitz and Ringertz (2001) point out, "it usually takes a longer time in economics (and social sciences in general) than in the natural sciences to find out if a new contribution is solid or if it is just a fad. In other words, it is important to wait for scrutiny, criticism and repeated tests of the quality and relevance of a contribution."

[10] The fact that a paper's evaluation may depend heavily on the characteristics of the evaluator is illustrated by the low correlation -only 0.24- that Blank (1991) found in the ratings given by the two referees of papers submitted to the American Economic Review.

mal evaluation will involve giving more weight to information that is observed along the dimension where they excel. This will favour candidates belonging to the same gender as the evaluator, as they are more likely to excel in that dimension. For instance, if the employer was male, papers produced by male candidates would tend to obtain a higher valuation since the (male) evaluator's optimal evaluation involves giving a higher consideration to the information observed across the mathematical dimension. Naturally, gender discrimination would disappear if the evaluator was well informed about the existence of such gender differences in the distribution of quality and, most importantly, was allowed to take candidates' gender into consideration.

2 The model

We build on the standard model of statistical discrimination presented by Phelps (1972) where an employer must select a candidate out of a pool of applicants in a context of imperfect information. Our main departure from the traditional framework is (1) to allow for the existence of multiple dimensions of ability and (2) to make the accuracy of the evaluation at each dimension depend on the evaluators' knowledge of this dimension.

Let us consider the case of an individual whose total quality depends on his abilities or skills in a number D of different dimensions or fields $[q_i = f(x_{i1}, ..., x_{iD})]$. These fields can be understood as different tasks that the worker needs to undertake or as separable skills that are required to perform a single task. Candidates' abilities are exogenously given and independently and normally distributed,

$$\mathbf{x}_i \rightarrow N\left(\mathbf{p}_i, \mathbf{\Sigma}\right)$$

where $\mathbf{x}_i$ represents the Dx1 vector of abilities, $\mathbf{p}_i$ is a Dx1 vector of mean abilities and $\mathbf{\Sigma}$ is the corresponding variance-covariance matrix.

Without loss of generality, we impose two simplifying assumptions on the populational distribution of quality. First, we restrict the variance of quality to be equal across dimensions and normalize it equal to one. With this constraint we want to avoid a more general case where ability may vary systematically more along certain dimensions.

$$Var\left(x_{id}\right) = 1 \qquad \forall d = 1, ..., D$$

Second, we assume that an individual's ability along a certain field is independent of his ability along any other dimension. In other words, the knowledge of an individual's ability in one dimension does not provide any information about his ability in any other dimension.[11]

$$E\left(x_{id}x_{id'}\right) = 0 \qquad \forall d \neq d'$$

[11] As long as there exists some kind of multidimensionality, this is, provided that quality in different dimensions is not perfectly correlated, dimensions could always be appropriately redefined such that this condition is satisfied.

These two assumptions restrict $\mathbf{\Sigma}$ to be a diagonal matrix where the on-the-diagonal elements are equal to one.

In this multidimensional framework let us consider the case where individuals' total productivity is not observable but evaluators can observe some noisy and imperfect signal of candidates' ability at each dimension. These signals could be interpreted as the result of some tests or job interviews and their value will be a function of the candidates' true ability at each specific field plus an error term η which is assumed to be independently and normally distributed with zero mean and finite variance.

$$y_{id} = x_{id} + \eta_{id} \qquad \text{where } \eta_{id} \rightarrow N\left(0, \sigma_{\eta_d}\right)$$

Moreover, let us assume that in each dimension the accuracy of the signal is independent of the quality of the candidate who is being evaluated.

$$E\left(x_{id}\eta_{id}\right) = 0$$

Given the above assumptions, evaluator h will infer the quality of candidate i in dimension d as the weighted sum of the signal observed in this dimension and the distributional prior, where the weight given to the signal will depend on how accurately this signal is perceived by the evaluator:

$$E_h\left(x_{id}/y_{id}\right) = \gamma_d^h y_{id} + \left(1 - \gamma_d^h\right) p_{id} \tag{1}$$

where $p_{id} = E\left(x_{id}\right)$ and $\gamma_d^h = \frac{E_h(x_{id}y_{id})}{E_h(y_{id}y_{id})} = \frac{1}{1+\sigma_{\eta_d^h}}$. If, for simplicity, we assume that a candidate's total productivity is equal to the sum of his quality at each dimension $\left[q_i = f\left(x_{i1}, ..., x_{iD}\right) = \sum\limits_{d \in D} x_{id}\right]$, it follows that:

$$E_h\left(q_i/y_{i1}, ..., y_{iD}\right) = \sum_{d \in D} \left[\gamma_d^h y_{id} + \left(1 - \gamma_d^h\right) p_{id}\right]$$

This is, employer h will take relatively more into account those signals that she observes in fields where she can assess information more accurately.

2.1 Similar-to-me-in-skills effect

Let us define an evaluation as being complex as the situation where an evaluator's relative ability to assess quality is positively related to her own quality. More precisely, in a context where, without loss of generality, D is equal to two, we define an evaluation as being complex if:

$$x_{h1} > x_{h2} \Longrightarrow \sigma_{\eta_1^h} < \sigma_{\eta_2^h} \tag{2}$$

It easily follows that when the evaluation is complex, an evaluator who makes an optimal use of the available information will give a larger weight to those signals that have been observed in that dimension where her own ability is larger. This is, given an evaluator h,

$$x_{h1} > x_{h2} \Longrightarrow \gamma_1^h > \gamma_2^h$$

As a result, faced with two equally productive candidates i and j, evaluator h will tend to give a higher evaluation to the candidate who excels in the same dimension where she herself is best. More precisely,

Proposition 1 *Similar-to-me-in-skills effect*

$$q_i = q_j,\ x_{h1} > x_{h2}\ \&\ x_{i1} > x_{j1} \Rightarrow E_h[q_i] > E_h[q_j]$$

Proof. The evaluator h, who observes at each dimension d a noisy signal of quality, can use y_d as a least-squares predictor of the candidate's true ability in that dimension, x_d, according to the regression-type relation $x_d = \gamma_d^h y_d + \left(1-\gamma_d^h\right) p_{id} + u_d$ where $E[y_d u_d] = 0$ and coefficient γ_d^h will be determined by the accuracy of the signal $\left[\sigma_{\eta_d^h}\right]$ [as in equation (1)]. The evaluator's expected valuation of candidate i's total productivity will be equal to:

$$E_h[q_i] = E_h[x_{i1} + x_{i2}] = E_h\left[\sum_{d=1,2}\left(\gamma_d^h y_{id} + \left(1-\gamma_d^h\right) p_d\right)\right] = \sum_{d=1,2}\left(\gamma_d^h x_{id} + \left(1-\gamma_d^h\right) p_d\right)$$

Candidate j's productivity can be estimated in a similar way. The difference in the expected observed quality of the two candidates is equal to:

$$E_h[q_i] - E_h[q_j] = \sum_{d=1,2}\left(\gamma_d^h x_{id} + \left(1-\gamma_d^h\right) p_d\right) - \sum_{d=1,2}\left(\gamma_d^h x_{jd} + \left(1-\gamma_d^h\right) p_d\right) = \sum_{d=1,2}\gamma_d^h\left(x_{id} - x_{jd}\right)$$

which is positive since $q_i = q_j \Longrightarrow x_{i1} - x_{j1} = x_{j2} - x_{i2} > 0$ and $x_{h1} > x_{h2} \Longrightarrow \gamma_1^h > \gamma_2^h$. ∎

2.2 In-group bias

As shown above, when productivity is multidimensional and the evaluation is complex, evaluators are more likely to give a higher evaluation to candidates alike to them. In this subsection we investigate whether the existence of this similar-to-me-in-skills effect can generate an in-group bias. This is, if individuals belonging to the same group tend to possess knowledge in the same dimensions, will evaluators have a tendency to prefer group mates over equally productive candidates from other groups? As we will see, it depends on whether the evaluator is well informed about the extent of inter-group differences and whether she is allowed to take them into account.

Consider that individuals may belong to different groups defined according to gender, age, or some other easily observable and exogenous characteristic. Let us also consider the case where there are only two groups g_1 and g_2 and where candidates' total productivity is independent of group belonging:

$$E[q_i / i \in g_1] = \overline{q}^{(g_1)} = \overline{q}^{(g_2)} = E[q_j / j \in g_2] \tag{3}$$

This assumption does not prevent the possibility that members of the two groups tend to excel in different dimensions. More particularly, let us represent the existence of group-related variations in the distribution of quality in the following way:

$$x_{id} = \sum_{g=g_1,g_2} x_{id}^{(g)} + \mu_{id} \qquad d = 1, 2.$$

where $x_{id}^{(g)} = \left(p_d^{(g)} + \varepsilon_{id}\right) c_i^{(g)}$ and $c_i^{(g)} = 1$ if candidate i belongs to group g and zero otherwise. Let us also assume that μ_{id} and ε_{id} are normally and independently distributed with zero mean. Therefore, $x_{id}^{(g)}$ measures the differences in dimension d that can be explained by the candidate's belonging to group g, and $p_d^{(g)}$ is the expected ability in dimension d of individuals in group g. Finally, let us, for simplicity, consider the case where the distribution of quality across groups is symmetric so that the following condition is satisfied:

$$p_1^{(g_1)} = p_2^{(g_2)} \ \& \ p_2^{(g_1)} = p_1^{(g_2)} \tag{4}$$

In this set up, evaluators will estimate candidates' quality in a similar way as in (1). Let us define $\lambda_{id} = \mu_{id} + \sum_{g=g_1,g_2} \varepsilon_{id} c_i^{(g)}$ and $z_{id} = \sum_{g=g_1,g_2} p_d^{(g)} c_i^{(g)}$. Therefore $x_{id} = z_{id} + \lambda_{id}$ and, given that $y_{id} = x_{id} + \eta_{id}$, it follows that in each dimension the relationship between quality and signal, net of the group effect, will be equal to $x_{id} - z_{id} = \gamma_d^h \left(y_{id} - z_{id}\right) + u_{id}$. Thus, $E_h\left(x_{id}\right) = E_h\left[\gamma_d^h y_{id} + \left(1 - \gamma_d^h\right) z_{id}\right]$ where $\gamma_d^h = \frac{Var(\lambda_{id})}{Var(\lambda_{id})+Var(\eta_{id})} = \frac{\sigma_\lambda}{\sigma_\lambda + \sigma_{\eta_d^h}}$.

In our analysis we will distinguish between two possible situations. First, we present the case where in their evaluation employers may take into account candidates' observable signals of quality but do not condition their evaluation on candidates' group belonging. Second, we study the case where the evaluators condition their evaluation both on the observed signals of quality and candidates' group belonging.

2.2.1 Non-discriminatory practices

Let us define as non-discriminatory practices those situations where evaluators do not condition their evaluation on candidates' group belonging $[\forall i \forall d \qquad E_h\left(z_{id}\right) = p_d]$. Several reasons may prevent evaluators from taking into account the group belonging of candidates. Evaluators may not be aware of the existence of differences in quality profiles across groups. As well, even if evaluators are well informed about these differences, they may be restricted not to use this information. This is the case, for instance, in many firms and institutions where the hiring process is subject to a strict equal opportunities policy.

Paradoxically, when members of different groups are, on average, equally productive but differ in their distribution of ability, if the evaluator does not or cannot take into account candidates' group belonging, individuals belonging to her own group will tend to be favoured.

Proposition 2 *Non-discriminatory practices yield discriminatory outcomes*

$$\begin{aligned}\overline{q}^{(g_1)} &= \overline{q}^{(g_2)},\, p_d^{(g_1)} \neq p_d^{(g_2)} \;\&\; E_h\left(z_{id}\right) = E_h\left(z_{jd}\right) = z_d \Longrightarrow \\ &\Longrightarrow E_h\left(q_i\right) > E_h\left(q_j\right) \qquad i, h \in g_1, j \in g_2, d = 1, 2.\end{aligned}$$

Proof. Without loss of generality let us assume that members of group g_1 tend to excel in dimension one $\left[p_1^{(g_1)} > p_2^{(g_1)}\right]$. Let us also for simplicity consider the case where the evaluator h is a typical group g_1 member such that $x_{h1} > x_{h2}$, so that from assumption (2) it follows that $\gamma_1^{(h)} > \gamma_2^{(h)}$. Then,

$$E_h\left(q_i\right) - E_h\left(q_j\right) = E_h\left[\sum_{d=1,2}\left(\gamma_d^h y_{id} + \left(1-\gamma_d^h\right) z_{id}\right)\right] - E_h\left[\sum_{d=1,2}\left(\gamma_d^h y_{jd} + \left(1-\gamma_d^h\right) z_{jd}\right)\right] =$$

$$= \left\{E_h\left(z_{id}\right) = E_h\left(z_{jd}\right) = p_d\right\} = \sum_{d=1,2}\left(\gamma_d^h p_d^{(g_1)} + \left(1-\gamma_d^h\right) p_d\right) - \sum_{d=1,2}\left(\gamma_d^h p_d^{(g_2)} + \left(1-\gamma_d^h\right) p_d\right) =$$

$$= \sum_{d=1,2}\left[\gamma_d^h\left(p_d^{(g_1)} - p_d^{(g_2)}\right)\right] = \{\text{by } (4)\} = (\gamma_1^{(h)} - \gamma_2^{(h)})(\bar{x}_1^{(g)} - \bar{x}_2^{(g)}) > 0$$

■

This is, in a framework where evaluating is complex, if groups differ in their distribution of quality and evaluators do not take into account group-belonging, they will assign a higher valuation to those candidates that excel in the same dimensions as they do and, since the distribution of ability across fields is group dependent, this bias will tend to favour candidates that belong to the same group as the evaluator. This is, not taking into account group priors is not only informationally suboptimal but, moreover, it generates discriminatory outcomes.

2.2.2 Discriminatory Practices

If employers observe that employees belonging to certain groups tend to perform better on certain dimensions, it is likely that these employers will update their beliefs and they will take into account this information in their evaluations, at least, as long as they are allowed to do so. If the evaluator can condition her evaluation both on the observed quality signals and on the group belonging of the candidates, then any two equally productive candidates will tend to obtain the same valuations independently of group belonging.

Proposition 3 *Discriminatory practices yield non-discriminatory outcomes*

$$\begin{aligned}\overline{q}^{(g_1)} &= \overline{q}^{(g_2)},\, p_d^{(g_1)} \neq p_d^{(g_2)},\, E_h\left(z_{id}\right) = p_d^{(g_1)} \;\&\; E_h\left(z_{jd}\right) = p_d^{(g_2)} \Longrightarrow \\ &\Longrightarrow E_h\left(q_i\right) > E_h\left(q_j\right) \qquad i, h \in g_1, j \in g_2, d = 1, 2.\end{aligned}$$

Proof. As in proposition (1), without loss of generality let us assume that members of group g_1 tend to excel in dimension one $\left[p_1^{(g_1)} > p_2^{(g_1)}\right]$ and let us also for simplicity consider the case where the evaluator h is a typical group g_1 member such that $x_{h1} > x_{h2}$, so that from assumption (2) it follows that $\gamma_1^{(h)} > \gamma_2^{(h)}$. Then,

$$E_h\left(q_i\right) - E_h\left(q_j\right) = E_h\left[\sum_{d=1,2}\left(\gamma_d^h y_{id} + \left(1-\gamma_d^h\right) z_{id}\right)\right] - E_h\left[\sum_{d=1,2}\left(\gamma_d^h y_{jd} + \left(1-\gamma_d^h\right) z_{jd}\right)\right] =$$

$$= \sum_{d=1,2}\left(\gamma_d^h p_d^{(g_1)} + \left(1-\gamma_d^h\right) p_d^{(g_1)}\right) - \sum_{d=1,2}\left(\gamma_d^h p_d^{(g_2)} + \left(1-\gamma_d^h\right) p_d^{(g_2)}\right) = \overline{q}^{(g_1)} - \overline{q}^{(g_2)} = 0$$

■

This is, if well-informed employers may condition their evaluation on the group belonging of candidates the outcome of evaluations will be independent of the evaluators' group belonging.

Let us summarize the empirical implications of the proposed model. First, the model predicts that when productivity is multidimensional and the evaluation is complex, evaluators will tend to take most into account information provided along those dimensions where they are more knowledgeable. As a result, evaluators will tend to give relatively higher valuations to those candidates who excel in the same dimensions as they do (Proposition 1). Second, the model predicts that if the different groups, although equally productive, tend to excel in different dimensions, an in-group bias may arise such that candidates belonging to the evaluators' group will tend to obtain higher valuations. In particular, this same-group effect will arise if the evaluator does not condition her evaluation on the candidates' group membership (Proposition 2). This might happen if the evaluator is not well informed about the existence of group differences or if she is not allowed to take candidates' group belonging into account. However, if a well-informed optimal evaluator is allowed to take into account group-belonging information in her evaluation, not only will the accuracy of the evaluation be higher but there will be no group bias (Proposition 3).

3 *The Weakest Link*

We test the validity of the model using data from a TV show: The Weakest Link . In this game, contestants must answer questions belonging to a number of different fields over a series of rounds. Then, after each round, each contestant is asked to declare which player was the weakest link. Thus, since candidates are themselves evaluators, it is possible to observe a proxy of their quality at each dimension and to study how this affects their assessments.

A fortunate feature specific to the Spanish version of The Weakest Link is the fact that questions are explicitly classified according to well defined categories. A computer selects randomly each subsequent question out of a database of 60,000 questions. Note that, since questions are randomly chosen, total quality can be considered as an additive function of quality across each dimension. Last but not least, using this set up has the advantage that in the game all evaluators observe the same information at the time of their decision and, most importantly, it guarantees that all evaluators share the same objective function, which is common knowledge. These are important features, as they allow to rule out the possibility that players differ in their assessments because they have access to different information or because they do not share the same objective function. Thus, it allows to discard a competing alternative hypothesis which also would explain why people tend to give more weight to that information which is observed in those dimensions where they are best at. In particular, if people differed in their assessment of which is the true importance of each dimension then individuals would invest more in human capital in those dimensions that they consider more important and, for a similar reason, they would also give a higher valuation to those candidates that excelled in those same dimensions.

Studies that use data from TV shows tend to be subject to several drawbacks. Individuals in the sample are not likely to be representative of the population. This sample bias is due both to individuals' self-selecting decision to apply to be in the show and from organizers' decisions regarding who is chosen.[12] As well, this program is broadcast on the national television and this may induce individuals to adopt politically correct attitudes, posing critical trouble for studies that attempt to identify discrimination. However, these caveats are less likely to affect the objective of our analysis: to estimate how an evaluators' relative knowledge of each field affects her assessments.

Below we describe the rules of the game, we present the information available in our database and then we analyze the factors that affect contestants performance and voting decisions.

3.1 The game

The game starts with nine contestants who have to work together as a team to build prize money of up to 7200 Euros. The show is divided in nine timed rounds. The first round lasts three minutes and each subsequent round is 10 seconds shorter. Within each round players take turns answering questions attempting to create a chain of correct answers. In particular, questions are asked to each player following a clockwise order until the time limit is exhausted.[13] A feature specific to the Spanish version of The Weakest Link is the fact that questions can belong twenty well defined categories.[14] Before hearing the question a player can make a banking decision in order to secure the amount of money in the chain.

[12] The show receives around 6000 applications each month, of which only 200 can be selected. One of the explicit criteria used to select participants for their show is their performance in a test trial.

[13] The first round starts with the contestant whose name is first by alphabetical order. From the second round on the best player of the previous round is the first to be questioned.

[14] Before asking any question the host indicates explicitly to which subject the question belongs.

If he does not bank and answers correctly, this increases the prize. Should he fail to give a correct answer, the chain falls back to zero. Immediately after each question is answered the host informs whether the answer was correct. A correct answer yields a link in a prize chain, beginning at zero and climbing to 800 Euros in nine increments. At the end of each round, players are asked to write in a blackboard who they think was the worst player in the round ("who is the weakest link?"). The voting is done simultaneously and then made public to all contestants. The player who gets the largest number of votes is eliminated from the game and leaves with no money. In the case of a tie, the strongest link decides which of the tied players is eliminated. By the end of round 7 only two contestants remain in the game and then play another round as a team to increase the final prize. In round 9 they compete head-to-head and the winner takes all the money accumulated during the show.

3.2 The data

Our dataset contains information on 103 episodes of The Weakest Link TV show that were broadcast in Spain between November 2002 and February 2004. The data were collected by video recording the episodes. In total, our data base covers information on the personal characteristics, the performance and the voting decisions corresponding to the 927 individuals who participated in the 103 programs recorded.

As it has been argued by previous works, while in the early stages of the game individuals' best strategy is to vote against players who they truly considered performed badly, as the game progresses, incentives may change as in the final round the two remaining contestants have to compete to get the prize (Levitt 2004, Février and Linnemer 2006). Also, the composition of the team in all rounds but the first will be endogenously determined by past voting decisions. Hence, to avoid these problems, our empirical analysis will consider only the first round of the show.[15]

3.2.1 Personal characteristics

Summary statistics at the contestant level are presented in Table 1.[16] Around 46% of players that participated in the show were female. This figure corresponds approximately to the calls and presentations to castings received.[17] The average age of contestants is around 36.8 years. In what follows we denominate younger/older players those below/above this age. According to this definition, in the game there around 29.1 % young females, 16.8% old females, 25.4% young males and 28.7% old males.

We also classify contestants according to their education level in five rough groups: less than high school, high school graduate, college graduate, student

[15] Results, available upon request, are essentially the same when we consider all rounds.

[16] At the beginning of the show players introduce themselves and generally report their age, their education and/or their occupation.

[17] Information provided in a personal conversation with the producer of the show.

and not known.[18] About 37% of the contestants are classified as college educated. With respect to the Spanish population, participants in the TV show are younger and better educated.[19]

3.2.2 Performance

The average player answers 59.1% of the questions correctly (see Table 2). Table 3 presents the distribution of contestants according to the number of questions asked and answered correctly. Most players are asked three questions (73%) and the mode was answering two out of these three questions correctly in the first round. Only around 19% of players manage to answer every question correctly and, on the contrary, only 7% do not manage to get a single question right.

Performance varies across different types of players (see Table 4). The rate of success is higher for men than for women (60.6% and 57.3%, respectively) and although this gap is not significantly different from zero at standard levels. Performance does not differ either significantly by age: the rate of success for players aged 36 or younger is 58.8%, very close to the 59.5% average obtained by contestants that are 37 or older. Within gender-age groups we observe again slight differences that are these differences are not statistically significant. Old female overcome younger ones (58.1% vs 56.8%) while, within males, young ones perform better (61.1% vs 60.2%). Significant differences are observed, though, across different educational groups. People that do not report neither their education nor their occupation perform the worst (51.8%), while people with college education tend to do best (62.2%).

We also observe significant differences in the difficulty of each field. While in fields such as Music or Literature players manage on average to answer correctly only 50.3% and 52.5% of the questions, the success rate in other dimensions such as Sports is 71%. Moreover, the relative knowledge of each field varies depending on the candidates' gender and age. Old men answer are relatively better than the average contestant at answering questions on subjects such as History (71% vs 54.4%), Politics (70% vs 62.9%) and Custom (73.1% vs 61.3%) and young men excel in Sports (79.7% vs 71%) and Entertainment (87.5% vs 64.7%). Among female, young ones are relatively better at Music (63.2% vs 53.9%) and Children (80.8% vs 72.7%) and old female stand out in fields such as at Gastronomy (68.4% vs 56.5%) or Religion (80% vs 60.2%).

4 What determines voting decisions?

At the end of each round players are asked who, according to them, was the weakest link of the round. As pointed out above, previous studies of The Weakest Link have shown that in the first round players' best response to this question

[18] In those cases where education was not reported, we have imputed the educational level that seemed most likely given the reported occupation. Whenever neither education neither occupation was reported players were assigned to the "not known" group.

[19] In Spain in 2004 the average age was 40.4 and around 28.2% of the population had attained tertiary education (OCDE, Education at a Glance, 2007).

is to truly declare who they consider was the worst contestant (Février and Linnemer 2006). However, evaluating contestants' ability presents several problems. First, the available information is scarce, as each player only gets to answer a few questions. Second, note that the first round lasts three minutes and on average 29 questions are answered, this is, one question is answered every six seconds. Given the speed of the game, it may be sometimes hard for players to remember other players' performance. Third, although the host reveals whether the question was answered correctly or not, not all questions provide the same information about an individual's quality, as not all the questions entail the same difficulty.

In fact, our data show that most players do not agree about who is the weakest link and, in general, votes tend to be spread. Only in five occasions out of the 103 sessions that compose our database the same contestant was voted as the weakest link by all the other players. On average, the most voted contestant gets slightly below five votes. The inexistence of a clear consensus raises the issue of how voting decisions are made, why voters tend to differ in their assessments of quality and whether group belonging plays a role.

Statistical discrimination models, as the one proposed above [see equation (1)], suggest that evaluators make assessments taking into account both the observed performance and their priors about candidates' quality, giving a weight to each piece of information proportional to its informative content. As well, contestants' personal characteristics may themselves influence voting decisions in the presence of taste discrimination.

We specify the utility that player h obtains from voting against player i as follows:

$$U_i^h = f_h\left(\mathbf{z}_i, \mathbf{y}_i, \mathbf{p}_i,\right) = \alpha + \beta \mathbf{z}_i + \frac{1}{N_i} \sum_{d \in D} \left[\gamma_d^h y_{id} + \left(1 - \gamma_d^h\right) p_{id}\right] + \epsilon_i^h \qquad (5)$$

where the set of covariates z_i captures contestant i's personal characteristics; N_i is the total number of questions that contestant i was asked; D is the total number of fields to which the question could belong; y_{id} represents the number of questions belonging to field d that individual i answered correctly; p_{id} reflects the evaluator's priors about contestant's i quality in dimension d; γ_d^h measures the weight that the evaluator gives to the observed signals and, finally, ϵ_i^h is the unobserved utility individual h obtains for voting against contestant i. Given this general specification, to carry out our analysis we estimate a conditional logit of the probability that contestants vote against other players. Under the assumption that the ϵ_i^h's are distributed i.i.d. extreme value, the probability of player h voting against contestant i is given by:

$$P_i^h = \frac{\exp\left[f_h\left(\mathbf{z}_i, \mathbf{y}_i, \mathbf{p}_i,\right)\right]}{\sum\limits_{i=1}^{C} \exp\left[f_h\left(\mathbf{z}_i, \mathbf{y}_i, \mathbf{p}_i,\right)\right]}$$

where C is the number of contestants in the round.

As Antonovics et al. (2005) point out, the advantage of using a conditional logit in this context is threefold. First, the characteristics and performance of all contestants will influence the probability that individual h votes against any one contestant. Second, the conditional logit allows us to examine interactions between the characteristics of the individual whose vote is being considered and the other contestants. Third, the predicted number of votes cast by each contestant is constrained to be one and therefore the total number of predicted votes cast in a round equals the number of contestants in the round.

4.1 Baseline model

Let us first assume a baseline case where (i) all questions are equally taken into account, independently of their field and independently of the characteristics of the observer $\left[\forall h, \forall d, \gamma_d^h = \gamma\right]$ and (ii) the priors held by evaluators do not vary across dimensions and individuals $[\forall i, \forall d, p_{id} = p]$. Our initial specification (5) is then reduced to:

$$U_i^h = a + bz_i + \gamma \sum_{d \in D} \frac{y_{id}}{N_i} + \epsilon_i^h \qquad (6)$$

This specification, which measures players' performance as the share of questions that have been answered correctly, is similar to the one used in previous studies that have analyzed voting behaviour in The Weakest Link (Levitt 2004, Antonovics et al. 2005).

Results from running regression (6) are reported in column 1 of Table 6. As personal characteristics we have included gender, age and educational level. We also control for the position where the contestant was located during the game and for the distance between players. In the game, contestants stand in a semicircle around the host. The physical distance between two players may be relevant, as a close distance could foster collusive behavior or increase the psychological cost of voting against each other. To measure distance, we number contestants from 1 to 9 and then compute the difference between these numbers. By construction, this variable varies between 1 and 8.

Confirming our expectations and also consistent with previous studies, a lower proportion of correct answers in the first round increases the probability of being voted out. Contestants are also more likely to vote against players who are located further away. Also, contestants who were asked a larger number of questions are more likely to receive a vote. This is consistent with an information-based theory, as a greater amount of questions answered diminishes the uncertainty about an individual's true quality (Cornell and Welch 1996). Also, as in Raghubir and Valenzuela (2006), we find evidence of a position effect.

While gender is not itself a relevant factor determining the voting decisions, older (younger) females have a greater (lower) probability of receiving a vote. This is consistent with the descriptive information presented in column 2 of table 4. Old women receive a significantly higher proportion of votes than the other three groups: old male, young male and young female. Given that older women's overall performance is not lower than that of young females, this evidence suggests that old females are treated relatively worse than young ones. Finally, higher levels of education are associated with a lower probability of receiving a vote, consistent with the idea that highly educated players are perceived as strong players even after controlling for performance.

These results differ partly from those found in previous studies that used data from other countries. For instance, Levitt (2004) observes that in the US older contestants are more likely to be voted but he finds no evidence of gender discrimination. On the other hand, Antonovics et al.(2005), using also American data, find evidence of female players discriminating against male ones. Finally, Février and Linnemer (2006) find, using French data, that gender and age do not have any influence on the probability of being voted out.

4.2 Similar-to-me-in-skills effect

In specification (6) the weight that evaluators give to signals, γ_d^h, is not allowed to vary across dimensions or across evaluators. However, Proposition 1 claims that, when the evaluation is complex an evaluator who makes an optimal use of the available information will give relatively more weight to signals that are obtained in those fields where she is relatively more knowledgeable. As a consequence, given two equally productive candidates, she will tend to give a higher valuation to the one who excels in the same dimensions as she does.

How complex are evaluations in The Weakest Link? The host reports whether an answer is correct, which facilitates players' evaluation. However, given that not all questions are equally difficult, a wrong answer may reveal bad quality or just reflect bad luck and, similarly, a correct answer may signal quality of the candidate or just be the consequence of good fortune. Therefore, being knowledgeable in a field might help to assess how informative an answer is about the quality of the player.[20]

Testing proposition 1 requires information on evaluators' relative quality at each field. It is possible to infer individuals' quality at each dimension by exploiting the information provided by players' performance during the game. In particular, we can proxy individual h's quality in dimension d using the share of questions that she has answered correctly in that dimension during the game:

$$\widehat{q}_{hd} = \frac{\sum\limits_{r=1}^{R_h} y_{hdr}}{\sum\limits_{r=1}^{R_h} n_{hdr}}$$

where r stands for round, R_h indicates the total number of rounds that player h participated in the game; y_{hdr} represents the total number of questions

[20] Antonovics et al. (2005) already argue that, in The Weakest Link, women (men) might be "better able to identify the types of questions that women (men) should be able to answer correctly". However, they model this using a standard (unidimensional) statistical discrimination model.

belonging to field d that contestant h answered correctly in round r; and n_{hd} indicates the number of questions belonging to field d that contestant h was asked in round r.

If we assume the existence of a linear relationship between an individual's knowledge and her evaluation accuracy, then an evaluator's relative accuracy at evaluating dimension d can be proxied by $\gamma_d^h = \frac{\widehat{q}_{hd}}{\sum\limits_{a\in D}\widehat{q}_{ha}}$.[21] Building on this measure, evaluator h will estimate individual i's performance as being equal to:

$$E_h\left(q_i\right) = \sum_{d\in D}\gamma_d^h y_{id} \tag{7}$$

We denominate this measure *observed performance*. Note that since $\sum\limits_{d\in D}\gamma_d^h = 1$, the *observed performance* is in expected value very similar to the share of questions answered correctly (see Table 2, rows 1 & 2).[22]

Additionally, we introduce as a control the type of questions that the player was asked. As expected, *observed performance* has a negative effect on the voting decisions (Table 6, column 2). In addition to the number of questions answered correctly, the empirical evidence confirms that individuals are less likely to cast votes against contestants with a similar knowledge profile. This is consistent with Proposition 1, which predicts that evaluators tend to prefer candidates that excel in the same dimensions where they excel.

Will the existence of a similar-to-me-in-skills effect yield an in group bias? In the next section we deal with this issue.

4.3 In-group bias

As it is shown in Table 5, there exist considerable differences in the distribution of quality across gender-age groups. Therefore, by construction, the variable *observed performance* should then be higher when both evaluator and candidate belong to the same group. The evidence supports this claim (see Table 7, column 1). The result also holds if we control for individual fixed effects (column 2).

[21] Given that quality at each dimension ranges between 0 and 1, one might consider also a Beta distribution with parameters a_d, b_d. With this specification, the weight that evaluator h gives to signal d would be equal to:

$$\gamma_d = \frac{\dfrac{a_d + \sum\limits_{r=1}^{R_h} y_{hdr}}{a_d + b_d + \sum\limits_{r=1}^{R_h} n_{hdr}}}{\left(\sum\limits_{d\in D}\dfrac{a_d + \sum\limits_{r=1}^{R_h} y_{hdr}}{a_d + b_d + \sum\limits_{r=1}^{R_h} n_{hdr}}\right)/D}$$

Results, available upon request, do not change if we use this alternative specification.

[22] Note also that as the *observed performance* has been built using the information provided by players' answers in the game, this measure will tend to be more precise for evaluators who remain longer in the show.

Propositions 2 and 3 show that when the distribution of ability differs across groups of candidates, whether the existence of a similar-to-me-in-skills effect generates an in-group bias depends on the following two conditions: (i) evaluators can condition their evaluation on candidates' group-belonging and (ii) evaluators are fully aware of the extent of intergroup differences.

If these two conditions are satisfied, our theoretical framework suggests that evaluators will take into account the available prior information about the existence of inter-group differences. This is informationally optimal and, moreover, in so doing players correct the in-group bias that would otherwise arise. Any available prior information about the relative quality of each group at each dimension should be considered, as it affects the information which is provided by a correct or a wrong answer. The intuition is easy: it does not provide much new information to observe an individual missing a question in a field which, anyway, members of his group tend to be unaware of. Similarly, it is not informative to observe an individual getting a right answer in a field where the evaluator already knows that members of his group are very knowledgeable.

While condition (i) is clearly satisfied in the game -players can observe the identity and group belonging of other participants- whether the evaluators are fully aware of the existence of significant differences in the distribution of quality across groups [condition (ii)] remains an empirical issue. In order to test if this condition is satisfied, we allow for a more general specification where priors are allowed to vary both across dimensions and across groups. In this case for any candidate i that belongs to group g, $p_{id} = p_{gd} = \frac{\sum\limits_{i \in g} y_{id}}{\sum\limits_{i \in g} n_{id}}$ where $\sum\limits_{i \in g} n_{id}$ denotes the number of questions belonging to field d that members of group g were asked and $\sum\limits_{i \in g} y_{id}$ measures how many of them they answered correctly. If we consider these priors, the best estimation evaluator h can do about player i's quality is given by:

$$E_h\left(q_i; p_{g1}, ..., p_{gd}\right) = \sum_{d \in D} [\gamma_d^h y_{id} + (1 - \gamma_d^h) p_{gd}] = \sum_{d \in D} \gamma_d^h \left(y_{id} - p_{gd}\right) + p_{gd} \quad (8)$$

which we name *observed performance with group priors.* Results from adding this variable to our original specification are presented in column 3 of Table 6. This measure has a negative and significant effect on the voting decision.[23] Furthermore, the Bayesian information criterion suggests that this specification of the model is better than the previous one.[24]

[23] The following anecdotical interpretation of this results might be done. If, for instance, a *young male* contestant fails to give a correct answer in a typically *old female* question (e.g. religion) this might be not considered as such a strong signal of low productivity as if the same mistake was made by an *old female* player. Hence, given two players that belong to two equally good groups and who obtained the same rate of correct answers, the one who answered correctly in fields where he was expected to be less knowledgeable is less likely to be voted out. Alternatively, the contestant who failed to answer questions in relatively (for his group) easy fields is more likely to receive a vote.

[24] The Bayesian information criterion of the model goes down from 3060.7 to 3058.5, indi-

These results suggest that players bear in mind the fact that the distribution of knowledge varies across gender and age groups and, consistently, adjust the information provided by signals. This means that the predictions of proposition 3 should apply. This is, when well informed evaluators take into account candidates' group belonging the existence of a similar-to-me-skills effect does not generate an in-group bias. In order to test this hypothesis we construct a dummy which takes value one if both evaluator and candidate belong to the same age and gender group and zero otherwise, and we add this dummy to the model specified in equation (6).

When we include this dummy in our baseline model, the coefficient is not significantly different from zero (see Table 6, column 4). This is, even if the people that look alike to the evaluator in terms of skills tend to be favoured and people within the same group tend to excel in the same dimensions, we do not observe an in-group bias.

5 Conclusion

In this paper we extend the standard model of statistical discrimination to a multidimensional framework where the accuracy of evaluators at a each field is related to how knowledgeable they are in that field. The model yields two main results. First, it rationalizes the existence of a similar-to-me-in-skills effect which favours candidates who excel in the same dimensions as the evaluator. The intuition is the following. Given that in this framework it is rational for employers to give more weight to signals observed in those fields where they are more skilled, candidates who stand out in these dimensions tend to obtain relatively higher evaluations. Second, the model casts doubts on the capability of blind evaluations to eradicate discrimination. It is shown that, if groups of individuals -according to their gender, race or any other observable and exogenously given characteristic- differ in their distribution of ability across dimensions, group discrimination will arise unless evaluators are well informed about the extent of these differences and, moreover, they can condition their assessments on candidates' group belonging. This is, when evaluations are multidimensional and complex, hiding the identity of candidates may actually penalize those candidates that do not belong to the evaluators' group.

We use data from the Spanish version of The Weakest Link TV show to test the main predictions of the model. As expected, we observe a similar-to-me-in-skills effect. In their voting decisions, even if all questions are equally valuable, players are more likely to vote against contestants who performed relatively worse in those fields where they themselves excel. We also observe the existence of significant differences in the knowledge profile of different gender-age groups. However, the similar-to-me-in-skills effect does not generate an in-group bias. Our analysis suggests that this is due to the fact that players are well informed about the existence of differences in the knowledge profile of different groups and take this information into account in their assessments. In this respect,

cating a better fit.

it provides exceptional evidence showing that the observability of candidates' identity might, in some cases, reduce discrimination.

References

[1] Aigner, D. J. and G. Cain (1977), “Statistical Theories of Discrimination in Labor Markets,” *Industrial and Labor Relations Review*, 30(2): 175-87.

[2] Aires, I. and P. Siegelman (1995) “Race and Gender Discrimination in Bargaining for a new car,” *American Economic Review*, 85: 304-321.

[3] Altonji, J.G. and R. M. Blank (1999) “Race and Gender in the Labor Market” in: O. Ashenfelter and D.Card. eds., Handbook of Labor Economics, vol. 3, (North-Holland, Amsterdam): 3143-3259.

[4] Antonovics, K., Arcidiacono, P. and R. Walsh (2005) “Games and Discrimination: Lessons from the Weakest Link,” *The Journal of Human Resources*, 40(4): 918-947.

[5] Antonovics, K., Arcidiacono, P. and R. Walsh (2008) “The Effects of Gender Interactions in the Lab and in the Field,” forthcoming in *Review of Economics and Statistics*.

[6] Arrow, K. J. “The Theory of Discrimination” (1973) in Orley Ashenfelter and Albert Reeds, eds., *Discrimination in Labor Markets*, Princeton, NJ: Princeton University Press: 3-33.

[7] Ayres, I. and P. Siegelman (1995) “Race and Gender Discrimination in Bargaining for a New Car,” *American Economic Review*, 85(3): 304-321.

[8] Bagues, M, Felgueroso, F. and M. Pérez-Villadóniga (2007) ”On the Optimal Composition of Evaluation Committees: Evidence from Public Exams,” mimeo, Universidad Carlos III.

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

[10] Bertrand, M. and S. Mullainathan (2003) ”Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination,” *American Economic Review*, 94(4): 991-1013.

[11] Blank, R. M. (1991) ”The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review,” *American Economic Review*, 81(5): 1041-1067.

[12] Byrne, D. (1971), ”The Attraction Paradigm*,”* New York: Academic Press.

[13] Chi, M. T. H. (1978). ”Knowledge structures and memory development.” (In R. Siegler (Ed.), Children’s thinking: What develops? (pp. 7396). Hillsdale, NJ: Erlbaum.)

[14] Chi, M. T. H., Glaser, R. and E. Rees (1982). Expertise in problem solving.(In R. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 1, pp. 1776). Hillsdale, NJ: Erlbaum.)

[15] Cohen, W. and D. Levinthal (1990) "Absorptive capacity: A new perspective on learning and innovation," *Administrative Science Quarterly*, 35:128-152.

[16] Cornell, B. and I. Welch (1996) "Culture, Information, and Screening Discrimination," *Journal of Political Economy*, 104(3): 542-71.

[17] Everson, H. T. and S. Tobias (1998) "The Ability to Estimate Knowledge and Performance in College: A Metacognitive Analysis," *Instructional Science*, 26: 65-79.

[18] Février, P. and L. Linnemer (2006) "Equilibrium selection: Payoff or risk dominance?: The case of the "Weakest Link," *Journal of Economic Behavior and Organization*, 60(2): 164-181.

[19] Fisher, M., Stanford B., Friedman M.D. and B. Strauss (1993) "The Effects of Blinding on Acceptance of Research Papers by Peer Review," *Journal of American Medical Association*, Vol. 272:143-146l.

[20] Fryer, R. G. and M. O. Jackson (2007) "A Categorical Model of Cognition and Biased Decision-Making," forthcoming in the *B.E. Press Journal of Theoretical Economics (Contributions).*

[21] Goldberg, Caren B. (2005), "Relational Demography and Similarity Attraction in Interview Assessments and Subsequent Offer Decisions," *Group and Organization Marketing*, Vol. 30(6): 597-624.

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

[23] Haan, M., Los, B. and Y. Riyanto (2004) "Signalling Strength? An Analysis of Economic Decision Making in The Weakest Link," mimeo, Groningen University.

[24] Holzer, H. and D. Neumark (2000) "Assessing Affirmative Action," *Journal of Economic Literature*, 38: 483-568.

[25] Hunter, W. C. and M. B. Walker (1996) "The Cultural Affinity Hypothesis and Mortgage Lending Decisions," *Journal of Real Estate, Finance and Economics* 13: 57-70.

[26] Kruger, J. and D. Dunning (1999) "Unskilled and Unaware if It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments," *Journal of Personality and Social Psychology,* 6: 1121-34.

[27] Lazear, E. P. and S. Rosen (1990) "Male-Female Wage Differentials in Job Ladders," *Journal of Labor Economics,* 8(1): S106-23.

[28] Levinovitz, A. W. and N. Ringertz (2001), ”The Nobel Prize: The first 100 years,” Imperial College Press and World Scientific Publishing Co. Pte. Ltd.

[29] Levitt, S. D. (2004) “Testing Theories of Discrimination: Evidence from The Weakest Link,” *Journal of Law and Economics*, 42(2): 431-451.

[30] Massey, D., and N. Denton (1993) “American Apartheid: Segregation and the Making of the Underclass,” Cambridge, Mass.: Harvard University Press.

[31] Neumark, D. (1999) “Wage differentials by sex and race:The roles of taste discrimination and labor market information,” *Industrial Relations: A Journal of Economy and Society*, 38(3): 414-45.

[32] Phelps, E. (1972) “The Statistical Theory of Racism and Sexism,” *American Economic Review,* 62(4): 659-61.

[33] Raghubir, P. and A. Valenzuela (2006) ”Center of Inattention: Position Biases in Decision Making,” *Organizational Behavior and Human Decision Processes*, 99(1): 66-80.

Appendix A
The Case of Early Arrivers and Late Leavers

Let us imagine a job where workers' total productivity (q) is just proportional to the number of hours worked and where the total amount of time worked by a worker is equal to the number of hours that elapse between her time of arrival (a) and her time of departure (d) $[q = k\,(d - a)\,; k > 0]$. Let us also assume that individuals cannot choose their arrival and departure times strategically, but these are fixed idiosyncratic characteristics, which are independently and uniformly distributed among the population $\left[a \rightarrow u\left(\underline{a}, \overline{a}\right) \text{ and } \rightarrow u\left(\underline{d}, \overline{d}\right), \text{where } \overline{a} < \underline{d}\right]$. For simplicity let us also consider that an individual's arrival and departure times are independent $[E\,(a \cdot d) = 0]$ and distributed with equal variance$\left[\overline{a} - \underline{a} = \overline{d} - \underline{d}\right]$. Finally, let us assume that the manager can only observe if an employee was working at a certain time if she was herself working at that very same moment. This is, if, for instance, the manager arrives at work at 10 a.m. and finds some employees already in the office, it is not possible for her to know exactly how long they have been there. Similarly, if the manager leaves the office at 6 p.m. and somebody else remains, she cannot know how much longer he will stay. In other words, in this example the accuracy with which the manager can evaluate the ability of employees at each "dimension" -arrival time, leaving time- depends on the manager's own "ability" in that dimension.

Similar-to-me-in-skills effect

Given any two employees with the same total productivity, the manager will tend to believe that the one who has a schedule closer to her own one is more productive. Let us first consider the extreme case where the manager's m type is such that she is the first to arrive in the morning $[a_m = \underline{a}]$ and the time of her departure will be at some moment between $\underline{d}$ and $\overline{d}$. Given that she is the first to arrive at work, she can tell perfectly how early all the employees arrive. However, with respect to the departure time, she will only know the exact departure time of those who leave earlier than her. For every employee who stays longer, given that she knows that arrival and departure times are independent, her best guess will be that any employee remaining will be staying until the mid point between the moment the manager leaves and the closing time $\left(\frac{d_m + \overline{d}}{2}\right)$. This is, given any employee i staying later than the manager, she will estimate his total productivity to be equal to $E_m\,(q_i) = E_m\,[k\,(d_i - a_i)] = k\,[E_m\,(d_i) - a_i] = k\left(\frac{d_m + \overline{d}}{2} - a_i\right)$, where the index m indicates the identity of the evaluator. Given any two equally productive employees i and j $(d_i - a_i = d_j - a_j)$ who leave later than the manager -*otherwise there would be error in the evaluation*- it easily follows that the manager will tend to give a higher evaluation to that employee whose profile is more similar to her own, in this case the one arriving the earliest in the morning $\left[a_i < a_j \Rightarrow q_i^{(m)} > q_j^{(m)}\right]$. A similar argument could be also developed for any type of manager.

Group discrimination

Now imagine that both managers and employees can belong to two different groups, Northerners and Southerners, and that group belonging is easily observable. Let us assume that on average members of both groups tend to work the same number of hours but, while people from the North tend both to arrive early at work and leave early, those from the South are more likely to arrive late and remain longer. In particular, let arrival and departure times be uniformly distributed such that

$$\forall i \epsilon G, a_i \rightarrow u\left(\underline{a}^G, \overline{a}^G\right) \ \& \ d_i \rightarrow u\left(\underline{d}^G, \overline{d}^G\right)$$

where $G = N, S$, $\underline{a}^N < \underline{a}^S$, $\overline{a}^N - \underline{a}^N = \overline{a}^S - \underline{a}^S$, $\underline{d}^N < \underline{d}^S$ and $\overline{d}^N - \underline{d}^N = \overline{d}^S - \underline{d}^S$

Following the above reasoning, if group-belonging was not observable, the manager will tend give a higher evaluation to that candidate who has a schedule closer to her own. Given two equally good employees, one Southerner and one Northerner, since people from the same group tend to have a closer schedule, the employee who belongs to the same group as the manager will tend to receive a higher evaluation.

Without loss of generality instance consider again the case of a manager who arrives earliest in the morning to the office and who, in this context, happens to be from the North. If she has to has to evaluate two equally productive workers, one from the North and one from the South, the observed expected productivity of each candidate will be given by $E_m\left(q_i\right) = k\left(\frac{d_m + \frac{\bar{d}^N + \bar{d}^S}{2}}{2} - a_i\right)$. Given that $E\left(a_i / i \in N\right) < E\left(a_i / i \in S\right)$, then $E_m\left(q_i / i \epsilon N\right) > E_m\left(q_i / i \epsilon S\right)$

This is, since the manager, who also is from the North, arrives early in the morning and can only observe the time of arrival but not of departure, she will tend to estimate that the most productive employee is the one who arrives early, who most likely will be the one belonging to her own group. This is, the existence of a similar-to-me-in-skills effect implies that when the evaluation cannot be conditioned on group belonging group discrimination will arise.

However, if the manager was aware of these group differences and she was allowed to take them into account then not only would her evaluation be more accurate but also group discrimination would disappear. The expected productivity of each candidate would be: $E_m\left(q_i / i \in G\right) = k\left(\frac{d_m + \overline{d}^G}{2} - a_i\right)$ where $G = \{N, S\}$.

Table 1: Contestants' characteristics

	Mean	Std. dev.	Minimun	Maximun
Female	0.460	0.499	0	1
Age	36.803	12.295	18	74
By gender-age				
Young female	0.291	0.455	0	1
Old female	0.168	0.374	0	1
Young male	0.254	0.435	0	1
Old male	0.287	0.453	0	1
By education				
Primary	0.190	0.392	0	1
Secondary	0.223	0.417	0	1
College	0.370	0.483	0	1
Student	0.141	0.349	0	1
Not known	0.076	0.264	0	1

Notes: The database contains information on 927 players. Young (old) refers to individuals who are 36 or younger (above 36).

Table 2: Performance

	Mean	Std. dev.	Minimum	Maximum
Percentage of correct answers	0.59	0.28	0	1
Observed performance	0.60	0.29	0	1
Observed performance with group priors	0.00	0.29	-0.80	0.57

Notes: The database contains information on 927 players. "Observed performance" and "Observed performance with group priors" have been computed as indicated in equations (7) and (8) of the text, respectively, which take into account information on both the voting and potential voted players. As each day there are nine contestants, each of them facing eight potential voting decisions, these variables have been calculated across the 7416 possible pairs of players.

Table 3: Distribution of contestants according to the number of questions asked and the number of questions answered correctly

		Number of questions answered correctly					
		0	1	2	3	4	Total
Number of questions received	2	4	5	2	-	-	11
		(0.43)	(0.54)	(0.22)	-	-	(1.19)
	3	52	196	293	136	-	677
		(5.61)	(21.14)	(31.61)	(14.67)	-	(73.03)
	4	10	34	73	82	40	239
		(1.08)	(3.67)	(7.87)	(8.85)	(4.31)	(25.78)
	Total	66	235	368	218	40	927
		(7.12)	(25.35)	(39.70)	(23.52)	(4.31)	(100)

Notes: Percentage values in parenthesis. The time assigned to the round is fixed and not all players receive the same number of questions.

Table 4: Performance and votes received by gender, age and education

	Correct answers (%)	Votes received	N
	(1)	(2)	(3)
By gender			
Female	0.573	1.099	426
	(0.014)	(0.085)	
Male	0.606	0.916	501
	(0.012)	(0.066)	
By age			
Young	0.588	0.891	505
	(0.013)	(0.070)	
Old	0.595	1.130	422
	(0.013)	(0.081)	
By gender-age			
Young female	0.568	0.893	270
	(0.018)	(0.095)	
Old female	0.581	1.455	156
	(0.024)	(0.160)	
Young male	0.610	0.889	235
	(0.019)	(0.102)	
Old male	0.602	0.940	266
	(0.016)	(0.087)	
By education			
Primary	0.591	1.273	176
	(0.021)	(0.145)	
Secondary	0.564	1.072	207
	(0.020)	(0.113)	
College	0.622	0.746	343
	(0.015)	(0.070)	
Student	0.588	0.855	131
	(0.025)	(0.141)	
Not known	0.518	1.614	70
	(0.034)	(0.228)	

Notes: Standard errors in parenthesis. ”% correct answers” represents the performance of contestants in the corresponding category in round 1. ”Votes received” is the average number of votes cast for players in the specified group in round 1. Each player receives, on average, one vote. Young (old) players are those who are 36 or younger (above 36).

Table 5: Performance by group and type of question

	Overall	Young female	Old female	Young male	Old male	Share
	(1)	(2)	(3)	(4)	(5)	(6)
Art	0.600	0.569	0 .541	0.667	0.632	0.061
	(0.036)	(0.066)	(0.083)	(0.083)	(0.064)	(0.004)
Children	0.727	0.808	0.467	0.875	0.652	0.029
	(0.048)	(0.079)	(0.133)	(0.069)	(0.102)	(0.003)
Cinema	0.503	0.485	0.394	0.565	0.563	0.059
	(0.038)	(0.062)	(0.086)	(0.074)	(0.089)	(0.004)
Communication	0.590	0.592	0.517	0.639	0.596	0.055
	(0.038)	(0.071)	(0.094)	(0.081)	(0.069)	(0.004)
Custom	0.613	0.571	0.667	0.444	0.731	0.031
	(0.051)	(0.095)	(0.105)	(0.121)	(0.089)	(0.003)
Entertainment	0.647	0.000*	0.600	0.875	0.500*	0.006
	(0.119)	-	(0.245)	(0.125)	(0.500)	(0.001)
Famous people	0.598	0.591	0.632	0.622	0.552	0.036
	(0.048)	(0.107)	(0.114)	(0.081)	(0.094)	(0.003)
Fashion	0.635	0.704	0.846	0.611	0.481	0.028
	(0.053)	(0.090)	(0.104)	(0.118)	(0.098)	(0.003)
Gastronomy	0.565	0.607	0.684	0.438	0.500	0.028
	(0.054)	(0.094)	(0.110)	(0.128)	(0.109)	(0.003)
Geography	0.609	0.585	0.548	0.702	0.605	0.076
	(0.032)	(0.062)	(0.078)	(0.067)	(0.056)	(0.005)
History	0.544	0.385	0.559	0.569	0.710	0.079
	(0.032)	(0.055)	(0.086)	(0.062)	(0.058)	(0.005)
Language	0.554	0.533	0.517	0.571	0.579	0.055
	(0.039)	(0.075)	(0.094)	(0.085)	(0.066)	(.004)
Literature	0.525	0.529	0.528	0.444	0.589	0.080
	(0.032)	(0.061)	(0.084)	(0.063)	(0.058)	(0.005)
Music	0.539	0.632	0.500	0.639	0.359	0.080
	(0.032)	(0.056)	(0.080)	(0.062)	(0.060)	(0.005)
Politics	0.629	0.588	0.692	0.550	0.700	0.023
	(0.058)	(0.123)	(0.133)	(0.114)	(0.105)	(0.003)
Religion	0.602	0.533	0.800	0.538	0.706	0.028
	(0.054)	(0.093)	(0.133)	(0.100)	(0.114)	(0.003)
Science	0.631	0.543	0.732	0.612	0.663	0.104
	(0.027)	(0.056)	(0.060)	(0.053)	(0.050)	(0.006)
Sports	0.710	0.636	0.633	0.797	0.707	0.080
	(0.029)	(0.065)	(0.089)	(0.047)	(0.051)	(0.005)
Television	0.353	0.500	1.000*	0.250	0.500	0.006
	(0.119)	(0.289)	-	(0.164)	(0.289)	(0.001)
Traveling	0.618	0.630	0.577	0.588	0.667	0.055
	(0.038)	(0.072)	(0.099)	(0.070)	(0.074)	(0.004)

Notes: standard errors in parenthesis. Figures in the table correspond to the 3009 questions that were asked in round 1. Young (Old) refers to players who are 36 or younger (above 36). An asterisk signals three cells with 2 or less observations. Not all of fields turn up with the same frequency. Share is the proportion of questions on each field.

Table 6: Voting behaviour: estimates of conditional logits

	(1)	(2)	(3)	(4)
Old female	0.397***	0.446***	0.471***	0.445***
	(0.130)	(0.137)	(0.138)	(0.138)
Young male	-0.011	-0.020	-0.024	-0.025
	(0.120)	(0.128)	(0.127)	(0.128)
Young female	-0.727***	-0.743***	-0.719***	-0.741***
	(0.172)	(0.181)	(0.181)	(0.183)
Secondary	-0.236*	-0.246*	-0.242*	-0.253*
	(0.124)	(0.129)	(0.129)	(0.129)
College	-0.322***	-0.289**	-0.283**	-0.301**
	(0.119)	(0.127)	(0.127)	(0.127)
Student	-0.320**	-0.248	-0.257**	-0.245
	(0.154)	(0.162)	(0.162)	(0.162)
Not known	0.062	0.060	0.044	0.061
	(0.160)	(0.170)	(0.169)	(0.169)
Distance	0.176***	0.180***	0.180***	0.177***
	(0.024)	(0.025)	(0.025)	(0.025)
% correct questions	-4.434***	-3.146***	-3.097***	-4.402***
	(0.172)	(0.594)	(0.530)	(0.178)
Observed performance		-1.246**		
		(0.572)		
Observed performance			-1.337**	
with group priors			(0.520)	
Same group				0.050
				(0.096)
Pseudo R2	0.2915	0.3053	0.3059	0.3040
Field Dummies	No	Yes	Yes	Yes
Number of observations	7416	7416	7416	7416

Notes: robust standard errors in parenthesis. In round one each player faces eight potential voting decisions, so there are 7416 possible pairs of potential voting-voted players. All regressions include dummies for position and for the number of questions received. Young (old) refers to players who are 36 or younger (above36). The omitted category is a male, older than 36 with primary education, who answered 2 questions and was located in position one. "Observed performance" and "Observed performance with group priors" are computed as indicated in equations (5) and (6), respectively.

*, **, *** indicate significance at 10%, 5% and 1%.

Table 7: Is observed performance higher when the evaluator and the candidate belong to the same age-gender group? OLS estimates

	(1)	(2)
% correct questions	0.997***	
	(0.002)	
Same group	0.005**	0.004**
	(0.002)	(0.002)
Individual dummies	No	Yes
Adjusted R^2	0.935	0.941
Number of observations	7416	7416

Notes: standard errors in parenthesis. The dependent variable is "observed performance". Both regressions include dummy variables for the evaluator. Same group is equal to 1 if both the evaluator and the candidate belong to the same age and gender group.

*, **, *** indicate significance at 10%, 5% and 1%.