LUND UNIVERSITY

DOUBTS ABOUT RATER OBJECTIVITY

An Investigation of Possible Ways to De-Bias Implicitly Biased Rankings

Julia Sjödahl

Master's Thesis (15 Credits)

April 13th, 2015

Supervisor: Martin Jönsson

Department of Philosophy

Lund University

TABLE OF CONTENT

ACKNOWLEDGMENTS

I would like to thank my supervisor Martin Jönsson for suggesting this topic for my thesis as well as for his guidance during the course of writing. I would further like to thank Cathrine Felix for comments on drafts of the thesis, and Jesper Sjödahl for proof-reading.

ABSTRACT

The aim of this paper is to present the kinds of error that implicit bias can cause when we are to judge and rank which available candidate is the most suited for e. g. a job vacancy, and to investigate whether we can de-bias ranking lists in recruiting processes given that we suspect that implicit bias has influenced the ranking. I will argue that the available methods for this (positive discrimination and the anonymizing of applicants, for example) are not sufficient, due to the fact that they do not take errors caused by implicit bias into account. Instead, I will investigate three possible ways to de-bias given the notion of a Borda score.

1. INTRODUCTION

Many professions are overrepresented by people belonging to specific social groups. Some professions are dominated by women, some by men, some by a specific ethnical group and some by people of a certain age – for example thirty and below or fifty and above. Some of this "skewedness" can be explained by people's career paths; it is, for example, hard to reach a position as a senior veterinary surgeon without having reached a certain age and without the corresponding work-related experience.

However, there seems to exist certain patterns that cannot be explained in this way. Why are there so few male elementary school teachers, so few female philosophy professors and so few black police officers? A possible explanation for this phenomena can be framed in terms of the notion of *implicit bias*; unconscious tendencies to judge and act in specific ways depending on what social group we are considering.

In this paper, I will focus on the negative aspect of implicit bias, i.e. when we judge or act in a way that yield negative consequences for the individual depending on his or her relevant social group. When hiring someone for a job, the outcome of the recruitment process might be affected by unconscious attitudes against women, men, black people, obese people, gay people, old people, young people, and so forth, unrelated to the individual's actual competence and suitability for the job.

As we will see, it is likely that we tend to systematically disfavour certain social groups in unjustified ways. From this, an epistemological problem arises: given that our judgements are affected by these tendencies, a lot of decision making in social situations is affected by completely irrelevant factors, and are, thus, likely made on the wrong grounds. The focus of this paper will be the question of how to prevent these irrelevant factors from affecting decision making in recruitment contexts.

I will begin with an explanation of the term implicit bias, and briefly describe some relevant empirical research on implicit bias in social psychology. In the subsequent sections, I will present the epistemological problem that implicit bias gives rise to, present some actual and tried de-biasing strategies connected to the recruitment process, and argue that these strategies are not sufficient. The final part of the paper consists of sketches of de-biasing strategies supposed to correct rankings from possible errors due to implicit bias.

2. IMPLICIT BIAS

2.1. THE TERM 'IMPLICIT BIAS'

Discussions of implicit bias in philosophy originate from findings within social psychology. Broadly, the term *implicit bias* is used by philosophers when referring to some sort of implicit cognition which, depending on someone's affiliation to a certain social group, affects our behaviour or judgements towards that person in a negative way through associating stereotypical traits with the person due to his or her affiliation to that group.[1] *Implicit cognition* can be understood as unconscious cognitive processes that are unavailable for one to report, but that to some extent affects one's judgements and actions.

Holroyd and Sweetman (forthcoming) argues that this broad sense, in which implicit bias often is defined, is problematic for two reasons. The first problem lie in the use of 'implicit bias' as referring to all sorts of implicit cognition, which could yield for unwarranted generalizations, and the second in tendencies to overlook the differences between the different processes often included. These tendencies are, according to Holroyd and Sweetman, reinforced by the tendencies to talk about the *effects* or *implications* of implicit bias. This implies that what we are talking about is a certain cause and its effects, rather than different processes falling under the rubric of implicit bias. Holroyd and Sweetman claims that these differences might be important when articulating normative advice about what to do about bias-related errors or injustice.

Holroyd and Sweetman do however agree that functional descriptions of implicit bias, such as the one I am using above, may be useful when considering questions about the *effects* of implicit bias. When one's priority is the effects rather than the causes of implicit bias, a functional description is useful since it makes it possible to distinguish distinct and useful patterns. This is exactly what I aim to do in this paper, which is why I will make no further assumptions about the metaphysical status of implicit bias.

[1] Not all agree in that implicit bias consists in mainly association. Some argue that they instead consist in propositional states, and others that they should be interpreted as 'cognitive schemas' (Brownstein, forthcoming). The nature of implicit biases is not a question I will address closer in this paper; its main focus will be on the implications they have on ranking processes.

2.2. IMPLICIT COGNITION AND MEASUREMENT

2.2.1. IMPLICIT COGNITION

Since the early nineties, *implicit associations* between concepts has been subject for studies within the psychology of implicit cognition. These associations obtains unconsciously all the time, relates all sorts of concepts, and may have positive as well as negative consequences. Sometimes, implicit associations obtains between a concept and an *implicit attitude*. Attitudes are beliefs, feelings or behavioural tendencies directed towards a specific concept. To sort out what implicit attitudes are, we need to distinguish them from *explicit attitudes*. Explicit attitudes are conscious and reportable, for example 'boys are wild', 'origin does not matter', 'women are forgetful' or 'teenagers are lazy'. Implicit attitudes, on the other hand, obtains unconsciously, but none the less affect our judgement and actions.

In order to illustrate implicit contra explicit attitudes, imagine Rhonda, a nurse, who explicitly holds the attitude that every patient at the hospital is entitled to equal care. What Rhonda does not know is that she carries an implicit attitude about women as more able to care for themselves than men, causing her to spend less time and care on a patient if the patient was female than she would if the patient were male. Hence, her implicit attitude is affecting both Rhonda's judgement (that female patients are in less need of her care) and her acting (spending less time and care on female patients).

Explicit attitudes are often consistent with our behaviour, but, as we can see in the case of Rhonda, this is not always the case. A striking example of this can be seen in experiments where the subject is asked to play a videogame where the subject is to shoot opponents carrying a gun. What these studies show is that an object is more likely to be interpreted as a gun in the hands of a young black man than in the hands of a young white man, and thus, the number of black men and white men killed in the game differ radically. Surely, not all the subjects in these kinds of tests hold the explicit attitude of themselves as more suspicious against black people than white.[2] Typically, the subjects are caught by surprise when they are informed that their test

[2] Sadly, what can be assumed to represent real life cases of so called shooter bias are easy to find. Examples are the recent killings in Ferguson (August 2014) and Cleveland (January 2015), where the Cleveland police shot a 12 year old African-American boy who held a toy gun.

results indicate implicit attitudes, since one expects from oneself to act according to one's explicit attitudes.[3]

2.2.2. EXPERIMENTAL RESULTS

2.2.2.1. CV STUDIES

Various studies has shown that the perceived social group of the name assigned to a specific CV impacts how the quality of that CV is perceived. These studies function in the following way: the subject is asked to judge the quality of a number of CVs. In the set of CVs to be judged, the very same CV appears, labelled with names usually associated with two different social group, so that the subject is asked to judge the same CV twice (or more), with the single variation that the name assigned to the CV differ.

A striking example is a study of this kind made by Marianne Bertrand and Sendhil Mullainathan (2004), responding to help-finding ads in Boston and Chicago newspaper. In this study, CV's with traditionally "white sounding" names got 50% more call-backs than CV's with "black sounding" names. Variations of this kind of experiment show similar results concerning various unprivileged groups; the social group of the applicant seems to be interfering with how we judge the quality of a CV.[4]

2.2.2.2. NEPOTISM AND SEXISM IN PEER REVIEW

In trying to sort out why so few women were granted postdoctoral fellowships from the Swedish Medical Council (MRC) in the early 90's, Wennerås and Wold (1997) investigated MCR's peer-review system. Wennerås and Wold's study resulted in highly disturbing conclusions about what factors affected which candidates were awarded fellowships.

In this case, the ranking functioned as following: the candidates were assesed by the board on three parameters: scientific competence, relevance of the research proposal and the quality of the proposed methodology. Each category were awarded 0-4 points. The applicant's points were then multiplied with one another so that each applicant could receive a score between 0 and 64 points, with the top scoring candidate coming out as winners. Wennerås and Wold noted that

[3] For more information, see Correll et. al. (2002, 2007), Greenwald, Oakes and Hoffman (2003), Payne (2001), Plant and Peruche (2005).

[4]See for example Moss- Racusin et al. (2012) and Steinpreis et al. (1999).

the MCR board awarded female applicants lower scores on all three parameters, resulting in an average score of 13.8 points for female applicants versus 17.0 points for male.

Wennerås and Wold measured the scientific productivity of each applicant in six different ways, (which I will present in section 2.3.3.). In one of the categories they measured, the aim was to find out each applicant's 'impact points'; a score supposed to measure each applicant's publications in scientific journals. It was shown that female applicants were rewarded lower rates than male applicants by the peer-reviewers. Wennerås and Wold arrived at the result that the most productive group of female applicants were in fact the only group of females (>100 impact points) who were considered as competent as the least productive group of male applicants (<20 impact points). Thus, we can see that the quality of the applicants work did not matter when it came to how the board judged them.

Wennerås and Wold found two additional factors which, except for scientific productivity, also highly affected the applicant's likeliness to receive a postdoc grant by MCR; namely gender and affiliation with a member of the board. For a female applicant to beat a male applicant she needed to exceed his scientific productivity with 64 impact points, or:

> "[…] approximately three extra papers in Nature or Science (impact factors 25 and 22, respectively), or 20 extra papers in a journal with an impact factor of around 3, which would be an excellent specialist journal such as Atherosclerosis, Gut, Infection and Immunity, Neuroscience or Radiology. Considering that the mean total impact of this cohort of applicants was 40 points, a female applicant had to be 2.5 times more productive than the average male applicant to receive the same competence score as he." (Wennerås and Wold, 1997, p. 3)

Being affiliated with one of the members of the board was shown to be worth 67 impact points. This results in that for applicants lacking personal contacts in members of the board to be viewed as equally competent as another applicant of the same gender associated with one of the board members, it was necessary to have a 67 point higher impact sore. Thus, a female applicant could make up for her gender by being affiliated with one of the board members. If not, she had to present an additional 131 impact points to win against a male applicant affiliated to a member of the board.[5]

[5] The use of Wennerås' and Wold's study is somewhat problematic since (1) the study is quite old, (2) there where quite few test subjects (114 applicants), and (3) the data seems to be lost (Sesardic & De Clercq, 2014), . I, however, think that a general description of the approach of Wennerås and Wold is useful since it reveals general structures in a ranking process supposed to be objective.

2.2.2.3. THE IMPLICIT ASSOCIATION TEST

The Implicit Association Test, developed by Greenwald and Banaji (1998) is designed to measure implicit associations by a subject's reaction time in classification tasks, and can be said to be the main instrument to expose implicit bias within subjects. Studies involving IAT has shown that almost everyone carry negative implicit attitudes towards certain social groups. Out of 700,000 participants, 70% where more likely to pair black faces with negative concepts and white with positive, indicating an implicit preference for white faces. In this study, most subjects answered the question whether they preferred black or white faces with that they had no preference (Nosek, 2007). A possible reason to why such a wide range of people has participated in the IAT is that it is available online at the Website of Harvard University for everyone who wishes to test themselves.[6]

The IAT consists in letting the subject sort concepts or pictures into different categories (for example: black/white, male/female, good/bad, and career/family) as quickly and with as few errors as possible. Subjects are supposed to be quicker to pair concepts that are closely linked together, indicating implicit association between certain concepts, which Nosek et al (2007) explains in the following way:

> "The IAT is a method for indirectly measuring the strengths of associations among concepts. The task requires sorting of stimulus exemplars from four concepts using just two response options, each of which is assigned to two of the four concepts. The logic of the IAT is that this sorting task should be easier when the two concepts that share a response are strongly associated than when they are weakly associated." (Nosek et al, 2007, p. 3)

When performing the test, the subject is placed in front of a screen and is in the first part of the test asked to pair, for example, one face with the more similar face out of two alternatives (example of categories of the two alternatives: a black face versus a white face, a thin face versus a fat face or a female face versus a male face). In the next part of the test, the subject is asked to pair concepts ("Love", "Unaccomplished", "Beauty" etc.) with the words "Good" or "Bad". After that, the lay-out of the test changes so that one of the alternative faces and the word "Good" can be seen on the one side (for example the white face paired with "Good"), with the other face and the word "Bad" (the black face paired with "Bad") on the other. Now, the subject is asked to choose the right side for the concept ("Love, "Unaccomplished", "Beauty" etc.) or face (black or white) appearing on the screen. The last part of the test is the same, but now the concepts "Good" and "Bad" are paired with the other face (so that "Good"

[6] The IAT can be found and taken at: https://implicit.harvard.edu/implicit/takeatest.html

is paired with the black face and "Bad" with the white). The subject's response time and error rate is assumed to show if the subject is carrying negative associations about a certain social group through the assumption that if the two concepts (white/good) share the same response, the response should be quicker than if they are weakly connected (black/good). [7]

The IAT's predicative force has been a subject of discussion. Nosek et al (2011) claim that implicit measure have unique predictive value, since it measure something different than explicit measures do, and that implicit and explicit measurement methods are predicative for different processes.[8] The IAT's predicative force is described below:

> "[..] Stronger implicit associations of self with death prospectively predicted suicidal ideation and actual attempt, implicit preferences for White people compared to Black people predicted voting for John McCain versus Barack Obama in the 2008 US presidential election, nations with stronger average implicit associations of science with males than with females have larger performance gaps favoring [SIC] men in science and math, and Swedish hiring managers' implicit racial bias predicted interview invitations for Swedish versus Arab-Muslim applicants. A meta-analysis of 184 predictive validity studies using the IAT found positive predictive validity across all evaluated domains" (Nosek et al, 2011, p. 155)

2.3. SOME AVAILABLE DE-BIASING TECHNIQUES

2.3.1. CAN WE USE THE IMPLICIT ASSOCIATION TEST?

The Implicit Association Test gives us an indication of which groups that most likely are subject for stigmatization and unprivileged treatment, by giving us an estimation of whether people tend to be biased towards a specific group or not. However, it seems complicated to apply this information in the process of de-biasing ranking lists. IAT is valuable in that it gives an indication of cases where bias is occurring, but not of how likely it is that one makes decisions on a biased basis.

2.3.2. POSITIVE DISCRIMINATION

A common strategy in attempts to produce *just* rankings is so called *positive discrimination*. The concept of this strategy is that the committee makes a ranking, and if the top two candidates differ in social category, the one from the unprivileged group is to be chosen. However, the obvious flaw in this method is that the implicit bias possibly affecting the ranking is ignored. If

[7] See Greenwald, McGee and Schwartz (1998).

[8] The IAT is, however, somewhat controversial for several reasons. It has for example been argued that it measures cultural knowledge and not strength of association (Arkes & Tetblock 2004).

we have a ranking list where the person from the unprivileged group does not reach the top two (possibly because of irrelevant bias-related factors), he or she will never be able to win. Another complication is that although this method may be ethically justified, it perhaps cannot be said to be epistemologically justified, since the applicant from the unprivileged group may not always excel the one from the privileged group.

Kang and Banaji (2006) however argue that given that candidates from different social groups have reached top two positions, we are generally justified to choose the one from the unprivileged group; since if the unprivileged candidate has reached the top two position despite the bias working against said candidate, he or she is very likely to be the strongest of the two. Whereas this is a good argument given the situation of top candidates representing different social groups, the possibility that the unprivileged applicant never reaches the top two due to bias related error appears unchanged.

2.3.3. WENNERÅS' AND WOLDS' METHOD

In their study about why women are less often granted postdoctoral fellowships by the Swedish Medicine Council, Wennerås and Wold (1997) investigated the female applicant's productivity in six different ways: total number of publications, total number of publications in which the applicant were the first author, the impact factor of the journals in which the applicants had published their work (judged from the number of times papers published in the relevant journal were cited in a year), the total impact factors of the journals were the applicant's papers were published, impact factors of the journals where first-author papers by the applicant were published, number of citations of the applicant's scientific papers during a year, and the same procedure for papers where the applicant was the first author.

Given these six parameters, Wennerås and Wold were able to count an 'impact score' for each participant depending on their scientific productivity, giving us a possibility to compare an applicant's score to other applicants' scores, and hence investigate if they were correctly placed in the ranking. One might be able to use this score as a way to generate rankings that are not influenced by implicit bias.

In cases where there is previous available scientific work, Wennerås' and Wold's method seems plausible. It also seems possible to transform this method into other domains than the scientific one as long as there is some sort of previous work by the candidates to judge. However, in cases where there are no, or perhaps very limited, earlier work available, this method does not work

quite as well, for example in cases where the applicants have not published any academic articles.

2.3.4. ANONYMIZING THE APPLICANTS

An obvious solution to many of the bias-related errors in rankings would be to anonymize the applicants and to make the decision without any knowledge of the applicants belonging to a given social group. Thus, the committee makes its decision unaware of the applicants' gender, age, ethnicity, and so forth. Saul (2013) exemplifies this through the fact that when orchestras began holding auditions behind screens, the number of female members increased dramatically. In many cases, this simple solution would solve the problem, but not in all. It would, for example, be practically limiting, since it would be impossible to interview the applicants in person. Another problem is that cases where the committee knows each applicant in advance are quite common. Disciplines where practically everyone knows everyone are not unusual.

Imagine a case where two people apply for a Ph. D. vacancy. They were both educated at the department in question, and the members of the committee supposed to judge the quality of the two candidates work are hence all familiar with it in advance (suppose that it is a small department). Among the committee members are even their former supervisor. Everyone, in the board know very well which student is associated with what papers, and it would in this case not matter if we erased the names on top of them; they would still not be anonymous, and hence, the decision of which candidate to choose would likely be made on wrong grounds.

3. BIAS-RELATED DOUBT AND RECRUITMENT

3.1. SCEPTICISM AND IMPLICIT BIAS

That implicit bias have moral implications might be quite obvious, since many of the situations described in the reported experiments have a direct application that involves injustice, but implicit bias also carries epistemological consequences. In *Scepticism and Implicit Bias* (2013), Jennifer Saul argues that the existence of implicit bias gives rise to a novel form of scepticism, *bias-related doubt*, and argues that what we know about implicit bias proves that we have reason to doubt a great deal of what we think we know.

Bias-related doubt is unlike traditional scepticism in that it does not consist of possibilities we are unable to rule out, but the fact that we are likely to systematically make errors on the basis of stereotypes related to social categories. It is a very serious form of doubt, and Saul puts it this way:

> "When assessing a contribution from someone who are [SIC] biases favour, we may grant more credibility than their testimony deserves; we may think their arguments are better than they are, perhaps failing to notice flaws that we would have noticed if the arguments were presented by someone else; we may take their evidence to be better than it is, and so on. And this is going to happen a great deal. It happens whenever we are dealing with the social world in a non-anonymised manner." (Saul, 2013 p. 251)

Saul argues that not only does implicit bias affect for example ranking processes and selection for academic journals, but it also has effects on what we hold as philosophical knowledge. Saul's main point is this: throughout misjudging the quality of a philosophical paper on a biased basis, we are also misjudging the quality of an argument since evaluation of the argument is unconsciously affected by irrelevant factors; namely the social affiliation of its author. Thus, implicit bias affects our entire range of philosophical beliefs, beliefs that we are *supposed* to have achieved through pure reason.

According to Saul, bias-related doubt is a form of scepticism that appears more serious than its traditional counterpart, since it is impossible to leave it in your armchair. It seems as if we cannot trust ourselves as inquirers. It is not only possible that we are making loads of decisions that are simply *wrong*; research on implicit bias suggests that we are doing it all of the time.[9]

[9] A common reply when faced with information about different kinds of biases is that these tendencies might very well exist, but that one oneself is unaffected by them. This is sometimes called "illusions of objectivity" – scepticism concerning one's own vulnerability of acting or judging in a biased manner. In a series of studies performed by David Armor (1999), 85% of the subjects estimated themselves as more objective than the average person. Participants were found to overestimate their own objectivity when faced with actual objectivity compared to other people in the same test group. The participants also tended to exhibit a naïve appreciation of the illusion of objectivity; they tended to attribute this phenomena to others, but not in their own judgements.

The data from the described experiments are supported by Pronin et al's (2004) argumentation that people in general are good at detecting biases in others, while denying that such biases might affects their own judgements and behaviour. One is often eager to retain the disposition of others response as different from one's own in a specific situation – so called naïve realism, which is argued to give rise to the belief that others are more likely to be affected by biases than ourselves. People tend to treat their own introspection of the basis of their judgements as superior to other people's introspective abilities, which is proposed to reinforce the conviction of one's own actions and judgements as non-biased.

A highly hypothetical consideration is that since philosophers (as well as mathematicians, physicists and so forth) are raised and trained to use specific objective methods when in one's work as well as when analysing other philosophers work and arguments, one might be slightly more vulnerable to the illusion of objectivity. Perhaps by coincidence, these are disciplines where quite few women are represented (in September 2011, as few as 24% of

3.2. THE AIM OF THIS PAPER

Since it seems to interfere with many social situation in our lives, it is possible to approach the problem of implicit bias from a number of different perspectives. A particularly interesting one is recruiting; Saul's argumentation indicates that the result of many recruiting processes is likely made on completely wrong grounds, and hence, in many cases simply incorrect.

As previously stated, there are cases where irrelevant and bias-related factors seems to affect which candidate is ranked the highest, and more specific; cases where there seems to exist some sort of non-spoken pattern (for example, one does not consciously aim at hiring only women in child care, but it seems to be happening anyway) in that specific social groups are seldom represented as winners in recruiting processes in a certain line of work. The core problem I am going to investigate in this paper can be defined thus: Given a group of applicants out of which at least one is member of a typically unprivileged group, how can we know which candidate best qualified for the position?[10]

In what follows, I will, hence, focus on the effects of implicit bias in recruitment processes, and not the many other factors affecting which applicant is chosen. One such other factor is the competence and previous experience of the applicants, another personal chemistry between, for example, the interviewer and the interviewee. Assume that the applicant and the interviewer get along great and accidentally have a mutual interest in horseback riding, have children in the

---

the permanent post-holders at UK philosophy departments were, for example, female (Beebee & Saul, 2011)). If this is the case, it would strengthen Saul's argumentation further.

[10] Here, a brief discussion about the status of possible confusion regarding social class-sorting in de-biasing processes might be useful. In most cases we can trust ourselves when it comes to identify class membership, that is, people are most commonly right when they, for example, identify someone with a typically male name as a man. There are, however, cases where the name does not give away any information of social group, or gives away the wrong information. An example of the first case would be gender neutral names, like Sam or Jamie, and an example of the second is parents who give their child a name with origin from another nationality than their own, or a name that is traditionally assigned to the other gender. This possibly causes errors in de-biasing, since we might make biased decisions grounded in mistakenly interpreted information. The apparent problem here is that people, also in de-biased rankings, might win or lose on wrong premises. I think that this may not be as big a problem as it may seem. If an applicant is disfavoured in a ranking because of his or her name, his or her placement in the list is likeably wrong independent of what social group he or she truly belongs to. The important thing about bias related error is that it occurs due to implicit associations we have, independent the applicant's individual characteristics.

same age or happen to live in the same neighbourhood. The difference is that personal chemistry may be a highly relevant factor when choosing the most suitable applicant for a certain vacancy, whereas the incorrect result that implicit bias cause on our judgements of an applicant's work should not be.

4. DE-BIASING: THREE POSSIBLE WAYS

Given the experimental results presented in section 2.2.2, I assume that general tendencies to favour certain social groups in different contexts exist. This is not to state that every person existing is automatically biased towards certain groups in specific contexts, but mainly to state that it is *likely* that one might have these tendencies. What I aim to investigate in the following parts of this paper are ways to de-bias final rankings in recruiting processes that we *suspect* to be affected by implicit bias, i.e. to present and investigate methods to find out if the rankings are or are not affected by implicit bias. If the ranking we are investigating and the de-biased ranking that comes out given the method are the same, no bias was operating, but if a difference appears it might be worth to take into consideration. A good de-biasing method should correct for implicit bias when it is occurring, but if it is not, a good de-biasing method should leave the ranking untouched.

4.1. BORDA COUNT

In my attempts to design a way to de-bias rankings, I will use the notion of a Borda Score. By doing this, each applicant will be provided with a score depending on where he or she has been placed in the ranking, which provides us with ways to manipulate possible ways to weigh up for the possible bias operating. It will be possible to do this for the individual's behalf, through his or her individual score, and for the unprivileged group's, through the aggregated score of the individuals within that group.

In a voting process using *Borda Count*, each voter ranks the candidates at hand in his or her preferred order. The candidates are then assigned a *Borda Score*. If the list consists of *n* (say 10) candidates, the candidate ranked first is assigned *n*-1 (9) points, the candidate ranked second *n*-2 (8) points and so forth, until we reach the lowest ranked candidate, who is assigned 0 points. The candidate who all together achieves the highest sum across lists when the voters' lists are accommodated wins the selection process (Pacuit, 2012).

So, if a board consisting of two members is to appoint a candidate for a job using Borda Count to reach a decision and there are four candidates, it could look like this:

Board member A's list:	*Board member B's list:*	
1. c = 3p	1. a = 3p	← *n*-1p
2. a = 2p	2. b = 2p	←*n*-2p
3. d = 1p	3. c = 1p	←*n*-3p
4. b = 0p	4. d = 0p	←*n*-4p

When combining these lists, we reach the following final list:

Final list

1. a = 5p
2. c = 4p
3. b = 2p
4. d = 1p

Thus, in this example, candidate a is the winner, since this candidate achieved the highest overall Borda Score.

My aim is to investigate a possible way to de-bias ranking lists by using Borda Count. I will focus on situations where we only have one ranking when attempting to work out a method that corrects for implicit bias, but the notion of a Borda Score is useful none to less, since it gives us a measure of how the candidates are ranked and provides a simple way to manipulate concerning their position in a ranking. My general purpose for using Borda Scores is that it provides a distinct way to manipulate with the applicants placement on the ranking in terms of a score. The aim of the presented methods is however somewhat wider; it would be possible to apply them to every ranking system assigning the candidate some sort of score.

4.2. THE PROBLEM

Imagine a situation where we are to choose the right person for a vacant position from a group of possible applicants. For the sake of the example, let's assume that we are looking for an architect to draw a building, and that we have received four applications. In this situation, we have applicants from two different specific social groups; men and women, and there is no real differences between the groups; the selection is made in an independent manner. We put the applicants through a process of interviews and tests, and in the end, the ranking comes out the following way:

a)

$n = 4$

1. John = 3p
2. Eric = 2p
3. Sarah = 1p
4. Jenny = 0p

Given the knowledge we have about men and women in ranking processes of this kind, we can assume that the group of men have bias-related advantages facing the group of women in this specific context (implicit bias concerning gender might very well work the other way around in other ranking contexts). So, we suspect an implicit bias to be interfering with this ranking, and ask ourselves what to do. What follows is an investigation of possible ways to update the list to correct from bias.

4.3. PRESENTATION OF THE METHODS

I will now in turn present and investigate three possible ways to go when searching for a way to de-bias ranking lists. The first one is based on the assumption that in a list where no bias is operating, the distribution of points between the social groups should be equal, the second is a improvement of that method, and the third is a way to calculate the degree of the bias and correct the list by the use of this number.

4.3.1. EQUAL DISTRIBUTION OF SCORES BETWEEN THE GROUPS

Consider the list presented in 4.2. In it, by assumption, two relevant social groups appear (we temporarily erase all other information that we may have about the people on the list); men and women. We can see that the group of men have received 5 points out of 6 and that the group of women have received 1 point out of 6 possible. A possible way to go to appears to be simply re-distributing the points equally between the groups so that the privileged and unprivileged group each receives 3 points.

A central property for an updated ranking is that it needs to keep the ranking within each group identical with the one in the actual ranking. A candidate ranked lower than its counterparts within the same group in the actual ranking should not possibly be ranked differently in the final, corrected, list. A suggestion of what the similarity between the original list ***a)*** and the updated one should consist in could be that the percent distribution within the group should be preserved in the updated list. Re-distributing the points equally, the possible updated lists appears to be the following:

a')	*a'')*	*a''')*	*a'''')*
1. John = 3p	1. John = 3p	1. Sarah = 3p	1. Sarah = 3p
2. Sarah = 2p	2. Jenny = 3p	2. John = 2p	2. Eric = 2p
3. Jenny = 1p	3. Sarah = 2p	3. Eric = 1p	3. John = 1p
4. Eric = 0p	4. Eric = 0p	4. Jenny = 0p	4. Jenny = 0p

In the original list, Sarah received 100% of the points within the group of women (1/1 point) and Jenny 0% (0/1 point), while in the group of men, John received 60% of the points (3/5 points) and Eric 40% (2/5 points). Hence, in the updated list where each group is assigned 3 points, Sarah should be assigned 3 points, John 2, Eric 1 and Jenny 0. Thus given, the list ***a''')*** is the correct list, and therefor also the list that should be used.

A problem with this specific method comes from the assumption that requested properties are supposed to be equally distributed within samples. The problem can be described as follows: imagine the property x to be equally distributed in the population. Even then, the property x would still be unequally distributed between men and women in samples of the population by random reasons. If one in one's de-biasing method assumes that x should be equally distributed in samples but only corrects rankings where for example women appears to have less of the property x than men, one creates injustice instead of preventing it.[11]

Another obvious problem with this de-biasing method is that in cases with very unequal distribution between the social groups, it yields for counter-intuitive results. Imagine a case with 99 applicants in social group *1* and only 1 in social group *2*. It seems counter-intuitive that the one member in *2* alone should be assigned the same amount of points that the 99 members in *1* are assigned to split; we risk to get updated lists extremely far from the actual list, but that would still count as correct, since the distribution of percentage would be preserved within each group. An example would be the following list, where 99 applicants has English sounding names and 1 has a foreign sounding:

b)

1. Carl = 99p
2. Linda = 98p
3. Robert = 97p

…

70. Margaret = 30p
71. Mohammad = 29p

[11] My supervisor Martin Jönsson pointed out this problem to me.

72. Diane = 28p

…

100. Samuel = 0p

Altogether, there are 5002 points to be distributed in list *b*, which means that the one member in group *2*, given the method we investigate, should be assigned 2501 points in every possible outcome of an updated list. This is, however, problematic, since the highest ranked candidate on the list can never be assigned more than 99 points. But, even if we do change the 2501 points Mohammed should be assigned into 99, he still beats every candidate in the group of English sounding name in every possible outcome. The updated list with the proper percent distribution would turn out as follows:

b')

1. Mohammad = 99p

2. Carl = 49.77p

3. Linda = 49.27p

…

70. Tom = 15p

71. Margaret = 14p

72. Diane = 13p

…

100. Samuel = 0p

Thus, the one candidate in group *2* still beats the top candidate in group *1* in every possible outcome, in this case simply because he or she is alone in his or her group. ***b')*** still is counter-intuitive and thus a problem for the method. It is clear that a proper de-biasing method needs to be sensitive to these kinds of situations.

4.3.2. PROPORTIONAL DISTRIBUTION OF SCORES BETWEEN THE GROUPS

Assume that the very same committee who ranked the list ***a)*** a year later finds themselves in a new recruiting process, and makes the following ranking of the four applicants:

c)

n = 4

1) Paul = 3p

2) Harry = 2p

3) Anna = 1p

4) Simon = 0p

In this list, 75% of the applicants are from the privileged group (men) and 25% from the unprivileged (women), and the points are distributed in such a manner that the privileged group has been assigned 83% of the points and the unprivileged 17%.

A possible way to go to de-bias given this information is to re-distribute the available points in accordance to it, so that the privileged group are assigned 4.5 points (75% of the available points) and the unprivileged 1.5 (25% of the available points). A central property for the updated list, as we saw in section 4.3.1, is that it needs to keep the ranking within each group identical with the one on the initial list. Thus, the points assigned to each group should in turn be re-distributed, so that each applicant keeps his or her percentage quota out of the total amount of points within his or her group. This list would come out in the following way:

c')

1. Paul = 2.25 (75% of the points within his group)
2. **Anna = 1.5 (100% of the points within her group)**
3. Harry = 1.49 (25% of the points within his group)
4. Simon = 0 (0% of the points within his group)

This way of de-biasing solves the problem of non-intuitiveness that appeared in section 4.3.1. If the unprivileged applicant is ranked 'to low' in the original list, he or she can never beat the top candidate from the privileged group, but with a sufficiently high original score he or she stands a chance. As we can see in example ***c'),*** Anna did advance, but she did not win.

The problem with this method, however, is that it does not seem to correct for implicit bias. Which candidate comes out as the winner is heavily dependent on the fact that the candidate is highly ranked in the initial ranking, as well as being member of the "right" social group. A lonely member in the unprivileged group facing a large privileged group will not win. On the opposite, he or she will achieve a very low score if we split the points in accordance to percent distribution between the groups in the original ranking. Which applicant comes out as the winning one depends on what the list looked like to start with, and the top candidate from the larger group will, given this method, therefore always win.

Another problem is that a possible outcome from this method is ending up with two or more applicants with the same final score. In these cases, we simply cannot epistemologically know which candidate to choose, but perhaps some ethical principle could guide us there. A possible, and perhaps reasonable, way to solve these situations would be to choose the applicant from

the unprivileged group in cases where two applicants from different social groups achieve the same score.

4.3.3. DEGREE OF BIAS

One of the key steps missing in the two methods presented so far, as well as in the de-biasing methods presented in section 2.2, is an estimation of the degree of the relevant bias. If we could know this, we would not face the difficulties with positive discrimination, nor risk that the unprivileged group would achieve too many points as in section 4.3.1.

Assume a method where the first and central step in the process of de-biasing a ranking is to find out the degree of the bias in the following way: Assume that the committee who ranked list ***a)*** have completed similar ranking processes attempting to choose the right person for a vacant position a number of times before, and that we have access to these past rankings.

Say, for example, that in all of the previous recruiting processes available to us, five applicants from a specific unprivileged group have applied for the vacancy. Given this information, we randomly select five applicants from the privileged group and five from the unprivileged in these lists, and accommodate a single list given these ten applicants (five from each group). In this sampled list, through dividing the Borda Score of the privileged group with the Borda Score of the unprivileged group, we get an estimate of the degree of the bias. A demonstration:

d)

1. Peter = 0.9p (9p) [12]
2. Daniel = 0.7p (7p)
3. Leonard = 0.7p (7p)
4. Lucy = 0.6p (6p)
5. James = 0.4p (4p)
6. Samuel = 0.3p (3p)
7. Sophie = 0.3p (3p)
8. Susan = 0.3p (3p)
9. Monica = 0.1p (1p)
10. Eleanor = 0.1p (1p)

[12] Since the scores come from different lists, they have been normalized. I will, however, in this list assume that every list we have sampled from consists of 10 applicants, why every score is divided by 10 (*n* of the list they are sampled from). Given lists with different number of candidates this serves to normalize their scores.

Men (privileged group) = 3p
Women (unprivileged group) = 1.4p

3/1.4 = 2.14

The degree of bias is **2.14**.

When the committee is to rank the applicants in a novel recruiting process, we, given the information of the bias degree in list ***d)***, de-bias the new ranking in a way I am now going to explain. First, the ranking of the applicants in the new recruiting process:

e)

1. Christopher = 9p
2. Walter = 8p
3. Tobias = 7p
4. Jim = 6p
5. Gerard = 5p
6. Matthew = 4p
7. Natalie = 3p
8. Rebecca = 2p
9. Arthur = 1p
10. Mark = 0p

What we do now is simply to multiply the unprivileged applicants' score in the current list with the degree of the bias from the sampled one. Doing this, we achieve the following updated list:

e')

1. Christopher = 9p
2. Walter = 8p
3. Tobias = 7p
4. Natalie = 6.42p
5. Jim = 6p
6. Gerard = 5p
7. Rebecca = 4.28p
8. Matthew = 4p
9. Arthur = 1p
10. Mark = 0p

As we can see, this method too has the property of taking the initial ranking into account, to avoid "extreme wins". Whether or not the unprivileged applicant is to win or not is simply determined by the degree of the bias. Extreme wins are only possible in contexts where the

degree of bias is very high and the ranking lists short. If an incompetent applicant from an unprivileged group is placed as one of the last in an original ranking of moderate length, it is theoretically possible, but highly unlikely, that he or she would reach the top positon.

Thus, a favourable consequence of this method is that if the unprivileged applicant is ranked 'to low' in the original list, he or she can never beat the top applicant in the privileged group, but with a sufficiently high original score he or she has the chance. It seems somewhat intuitive that one of the lower ranking applicants should not be able to beat the highest. However, this method does give a candidate of such rank the chance of advancing within the ranking in a way that seems accurate; for example, an applicant ranked second to last might very well advance, but perhaps not reach the top.

By providing us a way to actually measure the bias, it seem as if this solution in itself actually corrects for the bias in a way sensitive to its degree. The most obvious problem with the *Degree of Bias solution* is the question of how to de-bias rankings in scenarios where there are no previous ranking lists available. Differencing from the previously presented methods, this one requires that at least one person from the very same social group we are to investigate has been judged by the board at least once before. Of course, the issue here is that sometimes there are no previous applicants, or that it could be the first ever recruiting process made by a particular committee. This is an obvious weakness, but only given contexts where there are no previous lists. In many ranking processes there are in fact some kind of previous list, providing us with many situations where this method could be of practical use.

4.4. MULTIPLE BIASES

With the three methods presented, I will now investigate what possibilities they, each in turn, have when it comes to solving situations where two or more implicit bias might be working against a certain applicant. He or she might, for example, be both female and black or both old and disabled. In the following list, we assume that after holding interviews in a selection process for a financial manager, the committee is aware that the applicant Mary is in a wheelchair. We, hence, suspect that implicit bias might be working against her both because she is a woman applying for a job in a line of business where men are usually valued higher, and because she is disabled.

f)

1. Carl = 9p
2. Michael = 8p
3. Susannah = 7p
4. Richard = 6p
5. **Mary = 5p**
6. Iain = 4p
7. Michelle = 3p
8. Christian = 2p
9. Emily = 1p
10. Mathilda = 0p

A possible way to judge the quality of the three presented methods would be to investigate how good they are at handling multiple bias; something that can be assumed as constantly occurring in recruiting processes. What follows is such an investigation.

4.4.1. EQUAL DISTRIBUTION BETWEEN THE GROUPS

Given the method of Equal Distribution, we need to either investigate the two biases at hand separately and then combine the two lists, or assign the unprivileged applicant with 50% of the points, which we previously have found to be somewhat problematic. So, in list ***f)***, we have the bias against women in this specific context, and we have the bias against people in wheelchairs. Starting with the bias against women, we split the available 45 points in two, which gives each group 22.5 points. The 22.5 are then re-distributed according to the original percent distribution, giving us the following list:

f')

1. Susannah = 9.9p
2. **Mary = 7p**
3. Carl = 7p
4. Michael = 6.3p
5. Richard = 4.73
6. Michelle = 4.3p
7. Iain = 3.14
8. Christian = 1.58p
9. Emily = 1.23p
10. Mathilda = 0p

The next step is to investigate what the list would look like after fixing the wheelchair included bias. Since this is a case of unequally distributed groups (1 vs 9), we encounter the

same problem as we did in section 4.3.1 when splitting the points equally; as we can see Mary wins simply because she is alone in her group:

f'')

1. Mary = 22.5p
2. Carl = 5.18p
3. Michael = 4.5p
4. Susannah = 4.05p
5. Richard =3.38p
6. Iain = 2.25p
7. Michelle = 1.8p
8. Christian = 1.13p
9. Emily = 0.68p
10. Mathilda = 0p

Now, in accordance with the method, the two unprivileged groups have been given their correct amount of points in two distinct lists; the gender bias list and the disability bias list. In the first list, we got rid of the men's privilege over the women, and in the second, the non-disabled's privilege over the disabled. Now, we need to add a further step, which would be to combine the two lists into a final and correct ranking. A possible way would be to take the average of the two scores, which gives us the following formula:

Marys' score = (7 + 22.5) / 2 = 14.75p

Applying this formula on every applicant, we see the following result in the fully de-biased list:

f''')

1. Mary = 14.75p
2. Carl = 6.09p
3. Susannah = 7.2p
4. Michael = 5.4p
5. Richard = 4.06p
6. Michelle = 3.05p
7. Iain = 2.7p
8. Christian = 1.36p
9. Emily = 0.96p
10. Mathilda = 0p

When combining the lists, we do however encounter trouble. Firstly, in that we've added a further step to the method, and secondly in that any way we chose to combine the lists seems arbitrary. The presented way may seem close to our intuitions; we investigate and correct for the biases separately and then add the scores together to find out who wins. The problem with this is that if one of the biases we investigate have a stronger impact on our judgements than the other, say that a person in a wheelchair is more likely to be judged as an incompetent candidate for the job than a woman, maybe the wheelchair-bias list should be considered more heavily than the gender list. It seems as if we need to investigate the *women in wheelchairs-bias*, not simply the women-bias and the person in wheelchair-bias separately.

Another issue would be if there were two or more candidates with different combinations of multiple bias. What should the percentage distribution look like if one is not able to split the points in half between only two groups, but perhaps three or four? An example would be if for example Carl in this list is in a wheelchair as well or if Mathilda is black as well female. How should the points be distributed between the different social groups appearing when the method's way is to split the points equally between one privileged group and one unprivileged?

4.4.2. PROPORTIONAL DISTRIBUTION OF SCORES BETWEEN GROUPS

With the proportional distribution method, we also need to make to make separate investigations of each bias, and then combine the lists. We will start by calculating the correct list in respect to the gender bias. List ***f)*** provides us with the following information:

f)

Men = 29p (64% of the total amount of points)
Women = 16p (35% of the total amounts of points)

As we know, in list ***f)***, 50% of the applicants are men and 50% women, which gives each group an assigned score of 22.5 points. We redistribute the points in accordance to the original percentage score, and achieve the following result:

f’’’)

1. Susannah = 9.9p (44% of the points within her group)
2. Carl =6.98p (31% of the points within his group)
3. **Mary = 6.98p (31% of the points within her group)**
4. Michael = 6.3p (28% of the points within his group)
5. Richard = 4.73p (21% of the points within his group)
6. Michelle = 4.28p (19% of the points within her group)
7. Iain = 3.15p (14% of the points within his group)
8. Christian = 1.58p (7% of the points within his group)
9. Emily = 1.36p (6% of the points within her group)
10. Mathilda = 0 (0% of the points within her group)

The next step is to find out the correct list accounted for disability. List ***f)*** provides us with the following information:

f)

Non-disabled: 40p (88% of the total amount of points)
Disabled: 5p (11% of the total amount of points)

We know that 10% of the applicants belong to the unprivileged group and 90% to the privileged, and assign them their score accordingly:

f’’’’)

1. Carl = 8.91p (23% of the points within his group)
2. Michael = 8.1p (20% of the points within his group)
3. Susannah = 7.29p (18% of the points within her group)
4. Richard = 6.08p (15% of the points within his group)
5. **Mary = 4.5p (100% of the points within her group)**
6. Iain = 4.05p (10% of the points within his group)
7. Michelle = 3.24p (8% of the points within her group)
8. Christian = 2.03p (5% of the points within his group)
9. Emily = 1.22p (3% of the points within her group)
10. Mathilda = 0p (0% of the points within her group)

Now, with one list for each bias, we do however encounter exactly the same problem as the previously presented method did; the question of how we should combine the two lists. One of this methods strengths compared to the previous method is however its sensibility to percentage distribution in the original list, since it makes it possible to investigate as many social groups as required; there are no restrictions concerning what percentage of the points should be assigned each group more than the fact that they should be re-distributed in accordance to

percentage distribution in the original list. However, the problem, which was described in section 4.3.2, is that percent distribution does not corrects for an implicit bias.

4.4.3. DEGREE OF BIAS

With the Degree of Bias-method, it is easy to see how multiple bias are handled. We simply check the previous rankings for previous candidates from the same combined social groups and combine a list of randomly selected privileged candidates together, giving us the opportunity to investigate the exact bias, which previous methods were unable to do (by investigating each bias in turn). As mentioned earlier, the relevant bias in Mary's case would be the *women-in-wheelchairs-bias*.[13]

In the previous recruiting processes, we assume that there have been three applicants with the same combined social stigmas. Thus, the previous rankings it provides us with the following information:

e)

1. Paul = 0.93p (14/15)[14]
2. William = 0.64p (7/11p)
3. Laura = 0.4p (4/10)
4. Carla = 0.21p (3/14)
5. Tom = 0.12p (2/17)
6. Edith = 0.08p (1/12)

Men not in wheelchairs = 1.69p
Women in wheelchairs = 0.69p

11/4 = 2.45

The degree of bias is 2.45

[13] Here, it is important to be specific about the bias. We cannot simply say that the group of disabled women is the relevant group, since this group would have such a wide range and probably not capture the accurate bias. Perhaps we require a less wide description to describe Mary, something like *woman-who-is-in-a-wheelchair-because-of-a-spine-injury*. It is however an act of balancing, because at the same time we need the ability to look at the patterns working against women in wheelchair independent of reason.

[14] In brackets are the notion of the applicant's original score divided with *n* for that list.

Now, multiplying Marys' score with 2.45, the updated list turns out as follows:

f''''')

1. Mary = 12.25p
2. Carl = 9p
3. Michael = 8p
4. Susannah = 7p
5. Richard = 6p
6. Iain = 4p
7. Michelle = 3p
8. Christian = 2p
9. Emily = 1p
10. Mathilda = 0p

So, the degree of bias solution has (except for providing us a way to measure the bias) the advantage of, firstly, being able to investigate the biases side by side, and secondly by not having to add further, arbitrary, steps in order to de-bias.

5. RESULTS

The Degree of Bias method provides us with a way to measure the degree of the relevant implicit bias in a specific ranking, and is possible to use in every context where previous rankings are available. A strength of the method is that in many cases of recruitment, similar ranking processes have taken place previously. Thus, in many cases, provision of the previous rankings we need in order to apply our method would not be a problem.

A problem does however arise from the fact that previous rankings are required in order for this solution to be useful; sometimes there are no previous rankings available, which renders the method useless. A consideration on this matter is however that perhaps previous rankings are required for us to be able to know anything about the extent of a bias in a specific context. Since what bias is relevant seems to be contextually determined (gender bias might give slightly different results when hiring a nanny contra to when hiring a craftsman) we might need previous rankings in order to determine the bias in recruitment processes.

Another possible objection to this method is that it may seem too complex to use effectively in cases with a high number of applicants, as well as in cases where some applicants are sorted out before the actual ranking takes place. The first problem arises when a number of applicants

from a long list all seem perfectly equivalent, and consists mainly in the question of how to achieve a correct ranking when the quality of two or more of the applicants appears to be the same, i.e. whether or not the method is not fine-grained enough to handle this. If the candidates judged as equal were assigned different social groups, a possible way to go here as well would be to always rank the candidate from the unprivileged group the highest. This problem is however a problem for all kinds of ranking systems and not a matter I aim to solve in this paper. The other problem, I would say, could possibly be solved by applying the same method on the process of which candidates to sort out. If we need to cut out ten applicants before the proper ranking takes place, one could pick the ten less qualified candidates from each social group, make a ranking of them and apply the method, and thus achieving a de-biased list from which we keep the highest ranked candidates.

Perhaps some sort of solution based on the methods used by Wennerås and Wold (1997) could be useful as well. The demonstration of the existence of implicit bias throughout the Degree of Bias-method in a specific ranking process could possibly be strengthened by showing that the quality of the previous work of applicants belonging to a certain social group are misjudged on a biased basis, as well as the other way around.

6. CONCLUSION

I have investigated possible ways to de-bias rankings in recruitment processes. Beginning with an introduction to implicit bias and previous attempts to correct for social-group based patterns in recruiting, I continued on to a presentation of three different methods to correct for implicit bias: the Equal Distribution-solution, the Proportional Distribution of Scores Between Groups-solution and the Degree of Bias-method. The Degree of Bias-method seems to be the strongest one since it can handle multiple bias, preserves the initial ranking within the group, and lacks the problems the other two methods suffer from.

In ranking situations where we can observe specific patterns cause unjust chances for specific social groups when applying, we have tried to achieve unbiased rankings by methods such as positive discrimination as well as anonymizing the applicants. But as I have argued, these methods are not sufficient in all cases. Through the development of de-biasing methods, which do take implicit bias into account, bias-related errors in judging the competence of candidates would be reduced, causing the most competent and suited candidate to win. Thus, implementation of such a method would cause correct as well as just rankings.

If the Degree of Bias-method where to be implemented in recruiting processes, an expected critique of its applicability is its complexity; the Degree of Bias-method clearly contains more steps than current de-biasing methods commonly used. On the other hand, bias-related errors that easily slip through when, for example, using the positive discrimination-solution would not be able to happen; we can provide an actual number of the biasness within any given committee and change the ranking in accordance to that. The strength of the method is its complexity and ability to adjust itself to the biasness of each committee, providing a solution for that specific context. The Degree of Bias-method may seem somewhat complex, but so is the problem we are trying to solve.

7. REFERENCES

Arkes, H. R., & Tetlock, P. E. (2004) Attributions of Implicit Prejudice, or "Would Jesse Jackson 'Fail' the Implicit Association Test?" *Psychological Inquiry*. 15, 257-278

Armor, D. A. (1999). The Illusion of Objectivity: A Bias in the Perception of Freedom from Bias (dissertation)

Beebee, H. & Saul, J. (2011). Women in Philosophy in UK: A report by the British Philosophical Association and the Society for Women in Philosophy UK. Joint BPA/SWIP Committee for Women in Philosophy.

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

Brownstein, M. (forthcoming) Implicit Bias. *The Stanford Encyclopedia of Philosophy,* DRAFT 8.29.14

Correll, J., Park, B., Judd, C., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83, 1314–1329.

Correll, J., Park, B., Judd, C., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, 92, 1006–1023.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464-1480

Greenwald, A. G., Oakes, M. A. and Hoffman, H. (2003). Targets of discrimination: Effects of race on responses to weapons holders. *Journal of Experimental Social Psychology*, 39, 399–405.

Holroyd, J., Sweetman, J. (forthcoming). The Heterogeneity of Implicit Bias. In Michael Brownstein & Jennifer Saul (eds.), *Implicit Bias and Philosophy*. OUP.

Kang, J. & Banaji, M. R. (2006). Fair Measures: A Behavioral Realist Revision of "Affirmative Action". 94 *California Law Review* 1063

Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., Handelsman, J. (2012). Science Faculty's Subtle Gender Biases Favor Male Students. *PNAS* 109 (41) 16395-16396.

Nosek, B., Greenwald, A., and Banaji, M. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. *Automatic Processes in Social Thinking and Behavior*, in J.A Bargh (ed.), Philadelphia: Psychology Press.

Nosek, B. A., Carlee, B. H., Frazier, R. S. (2011) Implicit Social Cognition: From Measure to Mechanism. *Trends in Cognitive Science*. 15. 152-159

Pacuit, E, (2012). Voting Methods. *The Stanford Encyclopedia of Philosophy* (Winter 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2012/entries/voting-methods/>.

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlling processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, 81, 181–192.

Payne, B. K., Cheng, C. M., Govorun, O, & Stewart, B. D. (2005). An inkblot for attitudes: Affective misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3), 277-293

Peters, D. P., and Ceci, S. J. (1982) Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences* 5:187–255

Plant, E. A., and Peruche, B. M. (2005). The consequences of race for police officers' responses to criminal suspects. *Psychological Science*, 16, 180–183. Price-Waterhouse v. Hopkins, 109 S. Ct. 1775. (1989).

Pronin, E., Gilovich, T., Ross, L. (2004). Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self versus Others. *Psychological Review* 111 No. 3: 781-799

Saul, J. (2013). Scepticism and Implicit Bias. *Disputatio* 5: 37, 243-263.

Sesardic, N. & De Clercq, R. (2014). Women in Philosophy: Problems with the Discrimination Hypothesis. *Academic Questions* (vol. 27, no. 4)

Steinpreis, R., Anders, K., and Ritzke, D. (1999). The Impact of Gender on the Review of the Curricula Vitae of Job Applicants and Tenure Candidates: A National Empirical Study. *Sex Roles*, 41: 7/8, 509–528.

Wennerås, C., Wold, A. (1997). Nepotism and Sexism in Peer-Review, *Nature* Vol. 387/22