University of Reading

Bias and discrimination: what do we know?

Article

Accepted Version

Della Giusta, M. ORCID: https://orcid.org/0000-0003-3959-4451 and Bosworth, S. ORCID: https://orcid.org/0000-0002-8978-9516 (2020) Bias and discrimination: what do we know? Oxford Review of Economic Policy, 36 (4). pp. 925-943. ISSN 1460-2121 doi: 10.1093/oxrep/graa045 Available at https://centaur.reading.ac.uk/93414/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1093/oxrep/graa045

Publisher: Oxford University Press

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

www.reading.ac.uk/centaur

CentAUR

University of Reading

Central Archive at the University of Reading

Reading's research outputs online

Bias and discrimination: what do we know?

Draft Paper

Special Issue of the Oxford Review of Economic Policy "Gender Economics"

Marina Della Giusta and Steven Bosworth
University of Reading
20 April 2020

The paper presents the economic literature on gender bias, illustrating the underpinnings in the psychology of bias and stereotyping; the incorporation of these insights into current theoretical and empirical research in economics, and the literature on methods to contrast bias presenting evidence (where it exists) of their effectiveness. The second part of the paper presents results of an experiment in revealing unconscious bias.

Acknowledgments: we wish to thank Almudena Sevilla, Margaret Stevens and our second referee for thorough and insightful comments, and all participants in the Editorial Seminar for the Gender Issue of the Oxford Review of Economic Papers and in the seminar series of the Department of Economics at the University of Reading for very useful suggestions. We also wish to thank Sahira van de Wouw, who assisted with the Gorilla programming.

1. Discrimination in economics and in psychology

There is a vast empirical literature in economics documenting discrimination in many settings, from education and labour markets to access to credit, housing and health services, offers for products and services, politics, and law enforcement (Knowles et al., 2001; Bertrand and Mullainathan 2004; List, 2004; Nelson, 2009; Rodgers, 2009; Wood et al., 2009; Ewens et al. 2014; Alesina et al., 2018). Economists view discrimination as a situation in which individuals with identical productive characteristics are treated differently from each other on the basis of observable personal characteristics (such as age, ethnicity, gender, BMI, etc) that are unrelated to their productivity.

The tools economists have used to analyse discriminations have been mostly those of competitive markets in which discrimination arises from the behaviour of utility maximising individuals, and the efforts have been directed both at understanding why it may arise, why it may persist, as well as the consequences it has on individuals and society at large. The motivation for economist in studying the problem of discrimination have been thus both on grounds of equity and of efficiency: in regards to the latter standard theory would predict that a well-functioning market should gradually eliminate any selection based on non-economically relevant characteristics as it would be inefficient to discriminate, so theory has also sought to explain the persistence of discrimination in terms of imperfect markets (Stiglitz, 1973).

The seminal contribution in Gary Becker's Ph.D. dissertation (Becker, 1957) arose in a context in which discrimination against black and women workers in the US was legal (the Equal Pay Act dates to 1963 and the Civil Rights Act to 1964), and it was mostly the subject of

sociological rather than economic research, although it was the latter and especially the applied studies that were conducted on wage discrimination (Oaxaca, 1973) that was to have the strongest impact on actual litigation cases from then on (Ashenfelter and Oaxaca, 1987). Becker explained discrimination as taste-based, that is based on the dislike of a group on the part of discriminating individuals, who may be employees, co-workers or customers. In this case, discriminating employers would be effectively willing to pay the price of their choices by allowing a restriction in the pool of talent from which they select hires and promotions. In taste-based discrimination models those who dislike other groups are willing to pay a price to avoid interaction with them (Becker, 1957) and discrimination may thus persist even when it is inefficient. Over time, it may also persist because it becomes a social norm that is costly to break, as discussed by Akerlof in the context of explaining the causes of unemployment (Akerlof, 1980). The model of course assumes that individuals are rational and know their preferences, so that they deliberately discriminate in full knowledge of the costs this will have for themselves (if not those faced by those they discriminate). The solution in this case is for discriminated groups to become more skilled (thus costlier to exclude), so that discrimination becomes so expensive that even die-hard racists and sexists will be willing to forgo their exogenously given and perfectly known preferences in exchange for financial compensation.

The other classical explanation for discrimination in economics is that attributed to Arrow (1973) and Phelps (1972), who proposed statistical discrimination suggesting that when individuals have insufficient information about a particular person they will attribute to them average group perceived or real characteristics. These attributions amount to what psychologists describe as stereotypes, which are cognitive shortcuts used by the brain to generate expectations of other's behaviour and that can drive decisions based on unverified information about the group and individual belongs to rather than specific information about them, something Kahnemann and Tversky called the heuristics of representativeness (Kahnemann and Tversky, 1973). How these stereotypes affect belief formation has been the subject of work by both Baldiga Coffman (2014) and Bordalo et al. (2016). These models assume that significant biases in beliefs can arise from stereotypes or other conjunction fallacies (Kahnemann and Tversky, 1983) which consist in exaggerating small differences in some parts of the distribution of attributes of one group relative to another (for example believing that because men are over-represented in the top tail of the mathematics GPA scores distribution, this holds true across the whole distribution of scores). Economists have also modelled theoretically how the vicious cumulative effects of these decisions play out in aggregate outcomes: Oxoby (2014) has illustrated the process of forming beliefs about one's own ability incorporating irrelevant information on observable types can bias downward one's perception of one's own ability (or upward if the type-based biases are positive), and lead to inefficient allocations of agents across more and less skilled sectors in the labour market and a growing segregation over time through the feedback to agents from increased type-based biases in their beliefs. This type of discrimination can be corrected when information is supplied about the individual in question as these models expect rational individuals to correct their beliefs when they find them to be distorted. For economists, therefore, stereotyping based on group membership that results from imperfect information can be corrected through the provision of more information (Guryan and Charles, 2013).

Psychologists, though, have demonstrated that this is not so straightforward, and in particular stereotypes related to personal characteristics that are important for identity (as ethnicity and gender typically are) are not so easily corrected. In fact there is emerging neuroscientific research suggesting that cerebral networks used to process self-identity are different to those

used to process more general knowledge and much harder to change and thus correcting stereotypes with direct experience is difficult, particularly as other types of biases innate in the way we think come into play (Rippon, 2019): confirmation bias makes us pay much greater attention to the information that confirms what we already believe than to new information, and belief bias makes us forget information that are contrary to our beliefs (Kahneman, 2011).

Evidence of the length we are willing to go to protect our biases is provided in the recent paper by Bohren et al. (2019) who have conducted an experiment aiming to identify separate causes of discrimination by observing discrimination developing in a dynamic setting (a large online Q&A forum used by students and researchers in STEM). They have formally tested three hypotheses on the sources of discrimination: preference-based (a la Becker), belief based with correct beliefs (a la Arrow-Phelps), and belief-based with incorrect beliefs (a la Coffman). They find that decreasing subjectivity in judgment (through the provision of more information on the quality of the answers provided by female and male profiles) mitigates belief-based discrimination but does not affect taste-based discrimination which persists even when quality is perfectly observable. In particular, when discrimination is belief based (beliefs about a group's average ability) observing prior evaluation will reduce discrimination. Conversely, if discrimination is taste-based then even if they receive similar evaluation to men the women will continue to face discrimination. The mechanism they identify to explain this result runs through the different interpretations given to the signals received observing prior evaluations: in the case of biased beliefs evaluators become aware the woman had to produce work of higher standard to be positively evaluated, in the second instead the evaluator may believe she has been 'helped' and that the evaluations to not reflect her true quality (in other words they think the process is rigged). The paper shows both theoretically and empirically that reversals of beliefs can occur when discrimination is based on biased beliefs (stereotypes) which provides an important explanation for results observed in STEM where accomplished female academics are favoured over males (Williams and Ceci, 2015) and discrimination is instead found amongst female students (Reuben et al 2014), and in labour markets where discrimination occurs at hiring but reverses at promotion (Booth 2009, Groot and Van den Brink, 1996; Lewis, 1986), a result corroborated in the leadership literature that finds a gender premium at the top of organisations and discrimination at the bottom. The paper also provides evidence corroborating both Becker and contemporary psychology literature: when people are strongly prejudiced they would rather believe the system is rigged than change their beliefs.

The psychology of stereotyping (Jussim et al, 2015, Schenider, 2004) is fundamentally based on the fact that we are 'wired to be social' (Liebeman, 2013): our brains continuously work out extremely fast predictions on who to interact with and how, as described in the 'thinking fast and slow' model of decision making (Kahnemann, 2011). The fast mode unconsciously produces the statistics of repetition to recognise which groups of people we like (in-groups) and which do not like (out-groups) on the basis of stereotypes: shortcuts that incorporate a range of expectations of how someone else will behave. This is also true of self-expectations, that incorporate social rules of what is expected by someone like us. The slow mode conversely refers to deliberate thinking and works through problems more systematically, comparing and contrasting ideas and thus exercising more effort to arrive at more reliable decisions. Given the vast amount of decisions we make in everyday life, a lot of our interactions are driven by fast rather than slow thinking and incorporate the biases that it generates. A wide body of experimental evidence from both psychology and economics shows both that stereotyping and self/stereotyping are present, and that they can be artificially engineered in a wide variety of settings (Anderson et al, 2006). Whilst economists would like to think that system two is always

able to correct biases through belief updating, the evidence in psychology is cumulatively demonstrating that biases persist and even when revealed strong mean reversions occur in order to protect self-identity and reduce uncertainty.

Table 1: Concepts of discrimination

Economics We know what our tastes are and we know what our beliefs are We can fix biased beliefs through information provision, we cannot fix tastes	**Taste-based (Becker)** *You can show me women who are good at top maths, but I still won't hire them.*	**Correct beliefs-based OR Statistical (Arrow-Phelps)** *I think women are bad at top maths, and if you show me a woman who is good at maths I change my rating of her accordingly BUT keep my belief about women in general.*	**Incorrect belief-based (Coffman, Bordalo, Gennaioli)** *I think women are bad at maths in general, but would change my mind if I encounter many typically feminine women who are good at maths.*
Psychology We don't know our beliefs or our tastes very well We can somewhat mitigate biased beliefs through information	**Conscious bias (prejudiced)** Revealing unconscious bias activates immediate confirmation and belief bias that turns it into a conscious belief.	**Unconscious bias** Revealing unconscious bias activates short term re-evaluation and even attempts to over-compensate. Long run effects unknown (reversion to mean possible)	

2. Gender differences and gender bias

Gender is one important dimension along which stereotypes are formed (and interacts of course with other dimensions intersectionally), affecting behaviours and the directions of research itself (Rippon, 2019). Gender stereotyping includes both descriptive stereotypes, beliefs about what men and women typically do- derive from contact with each other (Fiske and Stevens, 1993), and prescriptive ones, beliefs about what men and women should do (Cialdini and Trost, 1998) which include positive proscriptions and negative ones (Koenig, 2018; Prentice and Carranza, 2002). For example women are supposed to be communal (warm, sensitive, cooperative; PPS for women) and avoid dominance (e.g., aggressive, intimidating, arrogant; NPS for women), and men are supposed to be agentic (assertive, competitive, independent; PPS for men) and avoid weakness (e.g., weak, insecure, emotional; NPS for men). The

psychological literature moreover tends to find that generally backlash is stronger for men and for boys transgressing the norms (Brown and Stone 2016; Sullivan et al 2018).

The evidence from the literature on psychological traits suggests that women on average are expected (both by men and by other women) to be more conscientious and compliant (Carter, 2014; Eswaran, 2014) and the self-reports that generate the data on personality traits (Big Five Inventory) shows women reporting on average higher levels of neuroticism, extraversion, agreeableness, and conscientiousness than men across most nations (Schmit et al., 2008; Costa et al., 2001). When it comes to evaluations of own ability, men on average perceive their general intellect as higher and they tend to overestimate it, whilst women on average tend to do the opposite (Karwowski et al., 2013)[1]. Goals reporting differs too: women on average declare that social objectives are more important than the goals connected with achievements, whilst men do the opposite (Kuhn and Villeval, 2015; Piirto, 1991). Recent findings from the Global Preference Survey (Falk et al., 2015) also suggest that women tend to exhibit a stronger social predisposition than men, and to be more responsive to social cues (Eckel and Fullbrunn, 2015; Zetland and Della Giusta, 2013).

Gender stereotypes emerge in early childhood and have important consequences (Bian et al, 2017). La Ferrara (2019) shows making use of PISA data across OECD countries that gender is strongly and robustly correlated with the probability of finishing university, with an effect that amounts in the most conservative specifications to a 15% increase over the mean. Worldwide the achievements and choices in maths by girls have been found to be strongly connected with the wider gender norms of societies (more gender equal societies have lower gaps- Guiso et al., 2008 in gender segregated schools girls choose STEM more and boys choose humanities more-Favara, 2012), with the gender of professors (Carrell, et al., 2010), teachers' gender views (teachers that have positive expectations increase the performance of pupils Figlio, 2005; Sprietsma, 2009; Campbell, 2015; Hannah and Linden, 2012. More gender egalitarian teachers increase the performance and uptake of STEM by girls- Alan et al., 2018; Carlana, 2019) parental beliefs (more gender equal parents have daughters that do better at maths - Cornwell and Mustard, 2013; González de San Román and De La Rica, 2012; Fryer and Levitt, 2010; Krings et al., 2014; Eccles, et al., 1990), and self-stereotyping (Coffman, 2014). Beyond the realm of human capital accumulation and its very real welfare consequences, Criado Perez (2019) has documented in her book *Invisible Women* the wide range of significant and persistent effects in the world of technology, medicine and politics arising from a design bias which takes men as the norm, and women as a deviation from it.

3. The effect of the exposure to stereotypes

Exposure to stereotypes effectively acts to hijack the brain by diverting resources to defending one's identity, activating a process of monitoring for failure and attempting to suppress negative thoughts both of which load working memory that would instead be needed to perform the task at hand (Schmader, 2010). This generates cognitive overload that diminishes performance. Poverty for example has also been shown to impede economic decision making (Shah, Mullainathan, and Shafir, 2012; Carvalho, Meier, and Wang, 2016; Adamkovič and Martončik, 2017) and performance on cognitively demanding tasks (Evans and Schamberg, 2009; Mani et al., 2013; Lichand and Mani, 2020).

[1] Parents also perceive their sons' intelligence to be higher than their daughters', while children perceive the intelligence of their fathers to be higher than that of their mothers (Karwowski et al., 2013).

Stereotype threat thus affects performance, and evidence suggests it can be manipulated for example to convince pupils to believe they belong to a group that has a natural advantage in a particular subject ahead of a test. More often than not it acts in negative ways: if pupils are reminded of their gender they do worse in the subjects in which they are expected to do badly, Mathematics for girls and English for boys (Johns et al., 2005; Jussim et al., 2015). The process starts very early: girls aged 4 have worse performance in spatial skills test if they colour in a girl playing with a doll before taking the test (Shenouda & Danovitch, 2014).

The literature has shown that exposure to bias toward one's group affects effort, self-confidence, and productivity (Carlana, 2019; Bordalo et al., 2016; Glover et al., 2017). Recent contributions in the development literature reviewed in La Ferrara (2019) explicitly link bias and stereotyping to the process of aspirations formation and aspirations as key contributing factors to poverty traps. Aspirations are strongly correlated to expectations (Carlana, La Ferrara, and Pinotti 2017), and expectations have been shown to affect performance, for example in the case of education independently of previous attainment and parental and other characteristics (Jacob and Wilder, 2010).

Exposure to counter-stereotypical role models has been shown to help improve maths performance by women when tests are administered by women whose competence is highlighted (Marx and Roman 2002). Breda et al (2018) have shown that exposure to female role models in schools can increase the proportion of female students who choose STEM subjects, and recently Porter and Serra (2019) have conducted an RCT showing that exposure to female role models majoring in economics has a positive effect on the choice of economics majors by women

4. Combating gender discrimination

Whilst much of the literature in economics has traditionally concentrated on assessing the effectiveness of specific policy interventions aiming to redress gender inequality (from equal pay legislation to parental leave or quotas) and more recently also on educational interventions aiming to increase confidence, grit, self-efficacy, leadership or math ability of girls (Alan et al, 2019 and 2020), there remains a large understudied area in establishing both the detection and the mitigation of discrimination arising from the psychological processes that feed into subjective elements of evaluation.

Discrimination often occurs along dimensions that are hard to quantify, such as the language used when engaging with and evaluating members of a targeted group. Combining evaluators from different backgrounds and making evaluations anonymous continues to be an important pillar of any strategy aiming to combat discrimination (Bohnet, 2016), but detection is anyway key to effective policy.

There is evidence that appraisal reports written for women are different than those written for men: Dutt et al. (2016) have investigated recommendation letters for postdoctoral fellowships in geoscience focussing on letter length and letter tone and found that women are significantly less likely to receive excellent recommendation letters than their male counterparts at a critical juncture in their career. Wu (2018) has measured gendered language in postings on the Economics Job Market Rumors (EJMR) forum and found that that the words most predictive of a post about a woman (*female* words) are generally about physical appearance or personal information, whereas those most predictive of a post about a man (*male* words) tend to focus on academic or professional characteristics. Bohren et al. (2019) found in their study a

significant difference in the sentiment of answers to questions from male versus female questions. Answers to questions posted by female accounts score significantly higher on both negative *and* positive sentiment; responses to female users contain more opinion words, both positive and negative, than responses to male users. The detection of bias in all these cases is likely to involve multiple dimensions, and research is still at the early stages, with algorithms offering promise (Kleinberg et al, 2019) though the latter can incorporate bias in their design or in fact might inadvertently exacerbate it (Schlesinger et al, 2018).

Unconscious bias training ("UBT") has been extensively adopted in many organisations, under the premise that by bringing our unconscious biases into conscious awareness, we can determine whether they are still appropriate behaviours and we can find ways to mitigate their impact on our behaviour and decisions (Equality Challenge Unit, 2013). Most UBT interventions include an implicit association test accompanied by a debrief, education on the psychological principles underpinning UB and its natural presence, information on its impacts and suggested techniques to mitigate it ranging from exposure to non-stereotypical situations to aid the removal of group characterisation in favour of individuation (Burns et al 2016 present evidence from South Africa), and the use of strategies to reduce the impact by requiring justification of choices, reducing the importance of subjective elements in the evaluation providing very clear performance indicators. There is evidence that appropriately designed UBT works for doctors, managers, and school pupils (Atewologun et al, 2018; Morris and Perry, 2017; Gilliam et al, 2016; Campbell, 2015; Devine et al, 2013; Teal et al, 2012). The principles of well-designed training have been identified as those that include an understanding of what stereotyping and unconscious bias are, that they have real effects and that we all have them and must accept this, that we must mitigate their effects creating a culture of recognition and promoting a culture of motivation of decision based as much as possible on a combination of clearly measurable indicators and bringing diverse perspectives to neutralize each other's biases. A recent meta-review of IAT based interventions (Fitzgerald et al 2019) found that whilst perspective taking was useful but only in the short run, exposure to counter stereotypical example was a quite effective tool.

The actual detection of unconscious bias in individuals has been widely done making use of Implicit Association Tests (https://implicit.harvard.edu/implicit/). The test typically requires to assign words to categories following both stereotypical (congruent) and non-stereotypical (incongruent) associations and measuring the speed with which such associations are made. For example to test for a gender and leadership implicit association one would have to assign female and male first names to words associated with being a leader or a follower in both congruent and incongruent scenarios.

Results of the test have been linked to negative expectations of employees (Reuben et al, 2014) and at the receiving end of the bias to the performance of employees of biased managers and pupil's performance and self confidence in math ability (Glover et al, 2017; Carlana, 2019). Revealing bias on its own can elicit subsequent moderating behaviours (Alesina et al, 2018) but is criticised by psychologists because of the possible negative reactions it can elicit (Howell et al, 2015) and most importantly because of likely subsequent effects that occur through moral licensing (Mazar and Zhong, 2010; Merritt et al 2012; Cascio and Plant, 2015), the process of behaving in a moral way but later being more likely to display behaviours that are immoral

which in the case of bias might lead subsequent to revelation, to reversion to previous biases beliefs and even backlash in subsequent longer term behaviours.[2]

It remains an open question in the social psychology literature whether exposing people to information about prior (im)moral behaviour leads to compensating or reinforcing subsequent behaviour. While the moral licensing effect has been replicated in many studies (see Blanken et al., 2015; Simbrunner and Schlegelmilch, 2017), Conway and Peetz (2012) point out that the moral licensing effect is at odds with results showing that people strive for consistency with past behaviour. Mullen and Monin (2016) argue that information is likely to elicit consistency when it is more abstract and temporally removed, when it is more relevant to the individual's identity, and when the individual's motives underlying their initial behaviour were ambiguous.[3]

Most of the literature however speaks to licensing effects arising from interventions to explicit pro- or anti-social behaviour. Monin and Miller (2001) show that allowing people to disavow overt (conscious) bias leads to a licensing effect. An evolving public discourse around unconscious bias has yielded a consensus that people may be morally responsible for their biased attitudes (Kelly and Roedder, 2008; Brownstein, 2015; Holroyd, 2015; Holroyd et al., 2017), especially to the extent that they are aware of them (Holroyd, 2015; Madva, 2017). Counteracting biased attitudes is seen as socially desirable, at least among certain population segments (Whiteout, 2018).

A crucial question then for practitioners seeking to counteract unconscious bias by making people aware of theirs is whether giving them an opportunity to improve (e.g. Carlana, 2019) is likely to generate licensing behaviour which counteracts the intervention or identity-affirming behaviour which reinforces it. In order to test whether bias revelation that gives rise to subsequent corrective behaviours also leads to further licensing, we have designed an experiment in bias revelation, the results of which are presented in section 5.

5. Revealing gender bias: an experiment

The aim of our experiment was to experimentally reveal to a selected group of participants their degree of bias using a common measure from the implicit association test (IAT) administered on a computer. These participants were then asked to repeat the IAT exercise on paper. This opportunity for redemption was intended to provide a moral license, which we then tested by asking participants to hand in their paper forms to two confederate researchers, one female and one male, whom they were told would tally the responses. This final behaviour (handing in the form) is our measure of moral licensing.

A total of 106 students completed participation in an undergraduate introductory microeconomics module at a large university in the South-East of England. The experimenters (also employees at this university) were introduced by the module convener during lecture time. Though participation was voluntary, most in attendance completed at least part of the study. The study consisted of two parts. In the first part, a link to an online Implicit Association Test (IAT) was sent to students' email addresses, and those with laptops present were asked to click on the link and complete it in class. The computer IAT was programmed in and hosted

[2] Bagues and Esteve-Volart (2010), Schwarz (2011) show for example that including more women on hiring panels can bias the selection process *more* towards male candidates.
[3] Clot et al. (2016) show an interaction between intended motive and social identity in the moral licensing effect: mandatory acts are more likely to elicit compensatory behaviour when the act is identity-relevant.

on the Gorilla.sc platform. The IAT first asked students to categorise a list of nouns as either Male or Female[4] and then a separate list of nouns as relating to Science or Liberal Arts.[5] Incorrect categorisations were communicated as such and participants were not permitted to move onto the next word until they assigned the word to the correct category. The two lists were then mixed in a random order and subjects had to allocate them to a combined category of 'Male or Science' or 'Female or Liberal Arts'. This condition is called the *Congruent* task since it conforms with gender stereotypes about STEM subjects being more male-oriented. A final task asked students to assign a re-mixing of these words to a 'Male or Liberal Arts' or 'Female or Science' category. This is the *Incongruent* task since it goes against the gender stereotype above. At the end of the computer IAT, half the students, randomly allocated by the server, were shown the difference in milliseconds of their average reaction times between the Incongruent and Congruent tasks. These subjects were also told that the difference in reaction times is used as a measure of their bias for associating science-based subjects with men. The other half of the participants were not shown their reaction times, but were similarly informed that their reaction times were recorded, and that this is a measure of psychological bias towards associating men with science-based subjects.

Following completion of the online IAT, a paper booklet was distributed to the students. This booklet was the same for all students and comprised a cover sheet on which students were asked for their student ID number and an instruction to wait for the experimenter before opening the booklet. Once all booklets had been handed out an experimenter instructed them to turn to the first page, which contained a vertical list of the above nouns, again in random order, with bubbles to fill in indicating whether this word was associated with a 'Science / Male' category or a 'Liberal-arts / Female' category. Participants were asked not to begin the task until the experimenter started a 30-second timer, and told that they should stop when the 30 seconds allocated were announced as over. The students were then told to turn to the next page, which again had the same list of words in a new randomised order, but the categories were reversed to be 'Liberal-arts / Male' and 'Science / Female'. Participants had 30 seconds again to complete the categorisation task.

Once all participants had finished, they were informed that we would use their student ID number to match their paper responses to their computer responses for comparison (individualised links had been sent by the platform and email addresses were associated with responses). We told students that once the matching had been done all identifiers would be stripped from the data for analysis, storage and reporting purposes. Finally, students were asked to hand in their paper booklets to one of two class tutors, one male and one female, who were positioned at the exit of the lecture theatre. Students were also informed that each tutor would tally the booklets handed into them.

5.1 Data

We recorded 123 complete responses to the computer IAT and 171 complete responses to the paper IAT. Of these, we were able to match 101 paper and computer responses.[6] This is because

[4] These words were Grandpa, Aunt, Son, Father, Sister, Wife, Girl, Woman, Mother, Man, Boy, Grandma, Husband, Uncle, Brother, and Daughter.

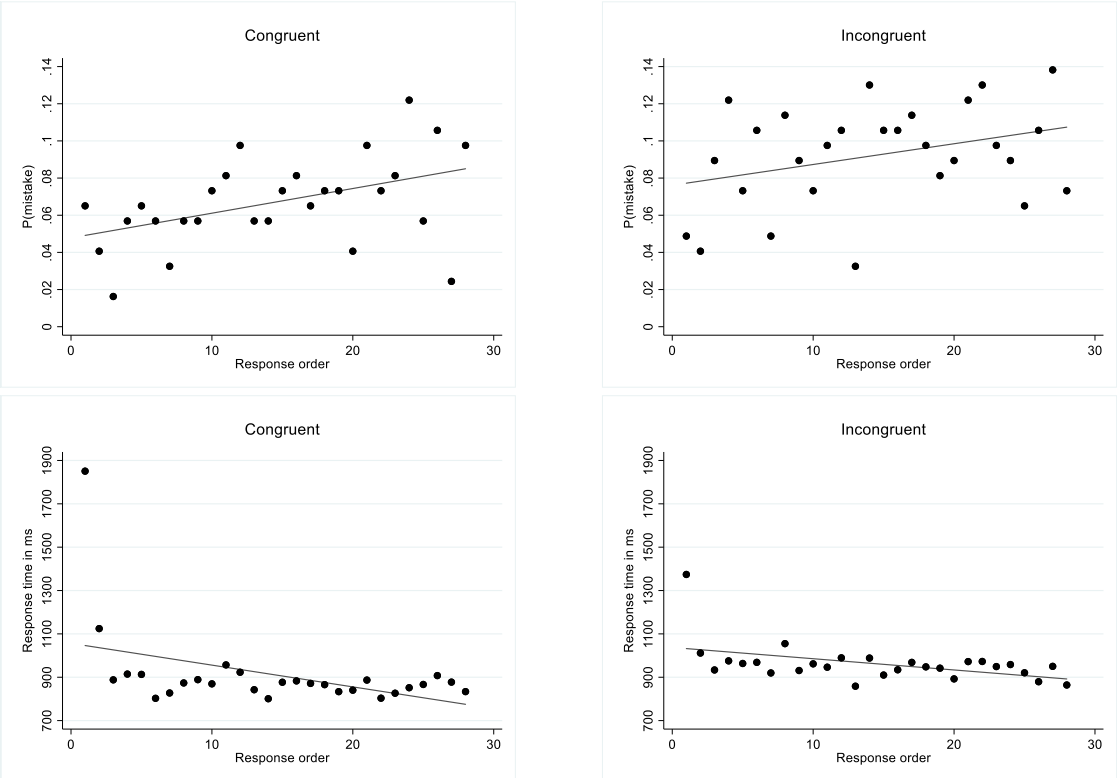
[5] These words were Literature, History, Geology, Music, Physics, Biology, English, Engineering, Astronomy, Humanities, Chemistry, and Philosophy.

[6] A further 8 subjects have paper IATs that can be matched to computerised IATs, but these subjects completed the computer IAT *after* the classroom experiment had finished and therefore their paper IAT responses are not commensurate with the others. These subjects are not included in the analysis.

some students did not complete the computer IAT and some students either did not complete the paper IAT or did not record their student ID number on their paper booklet. We report results only from the 101 matched cases. Among the matched sample 48 participants received feedback on their computer IAT and 53 subjects did not receive feedback. 74 participants were male whereas only 27 were female; this is unbalanced but is reflective of the gender balance in the economics degree programmes at this university, which are approximately 2/3 male. 50 participants handed in their paper IAT to the female tutor and 51 to the male tutor.

While the computerised version of the IAT enforces that subjects eventually choose the correct response to the categorisation question, subjects were able to make any number of mistakes before selecting the correct of the two responses. The average number of mistakes per response and average total reaction time per response are shown in Figure 1. Note that under both the congruent and incongruent conditions subjects were given 28 words to categorise. The horizontal axis represents the order they progressed through these words, which could have been different for every subject. We see that as subjects progress through the computerised task they gradually make more mistakes as they answer the questions slightly faster. The first word presented under each condition takes markedly longer for subjects to categorise than all the others, possibly due to attention lags. Unconscious bias on computerised IATs is most frequently measured though reaction times, incorporating the time taken for mistakes, so we use these going forward.

Fig. 1: Average probability of mistakes (top) and total reaction times for items (bottom) across congruent (left) and incongruent (right) conditions. Horizontal axis is the order in which each item is presented



Participants' total reaction times to both the Congruent and Incongruent computer tasks were recorded in milliseconds. From these numbers we compute a bias score in the computerised task, $bias_{Ci}$ for subject i by

$$bias_{Ci} = \frac{\overline{incongruent\ task\ reaction\ time}_i - \overline{congruent\ task\ reaction\ time}_i}{\sqrt{\sigma^2_{incongruent,i} + \sigma^2_{congruent,i}}},$$

dividing the subject's mean difference in reaction times between the congruent and incongruent conditions by their standard deviation of reaction time to control for heterogeneity among participants in ability and task engagement.[7]

From the paper IATs, Recall that subjects were under time pressure for the paper IATs and reaction times aren't observed, so our primary measure of cognitive association here are correct answers. [8] Subjects' total number of correct categorisations in both the Congruent and Incongruent tasks were recorded and from these numbers we compute a bias score in the paper task, $bias_P$ by

$$\begin{aligned} bias_P = (&congruent\ correct\ categorisations \\ &- incongruent\ correct\ categorisations) \\ &\times \left(\frac{congruent\ correct\ categorisations}{incongruent\ correct\ categorisations}\right)^2, \end{aligned}$$

again to control for any participant heterogeneity unrelated to gender bias. The bias scores were then made comparable by ranking them from lowest to highest among participants and computing percentile scores $\%bias_C$ and $\%bias_P$. Relative improvement between the computer and paper IAT tasks was taken as the difference in percentiles:

$$improvement = (\%bias_C - \%bias_P)/2,$$

with $improvement > 0$ corresponding to those who moved down in the bias ranking between computer and paper IATs and those with $improvement < 0$ moving up in the bias rankings. This was rather variable, with the minimum improvement being negative 37 percentiles and the maximum being positive 45 percentiles, with a standard deviation of 19.5. By construction mean improvement was zero.

5 Results

Figure 2 shows the percentile bias scores across the Feedback and No Feedback treatments among women (left panel) as well as men (right panel). Firstly let us remark that bias is similar among both men and women. There are no significant differences in bias rank across genders for either the computer or paper tests. Secondly, participants exposed to feedback about their gender bias are very slightly more likely to improve in the bias ranking when completing the

[7] Greenwald, Nosek, and Banaji (2003) suggest this method for scoring computerised IATs as preferable.
[8] This is suggested by Lemm et al. (2008) as the appropriate analogue to reaction time in paper IATs. We also adopt in calculating $bias_P$ the scoring method they suggest as having the highest correlation with Greenwald et al.'s computerised IAT score.

paper test than those who do not receive feedback, and this is not statistically significant (ranksum p = .45).

Figure 3 plots the relationship between the percentiles of both computerised and paper IATs by feedback condition for both women and men. There is surprisingly low correlation between the two IAT measures, likely due to the information presented to subjects that this was a measure of gender bias. The correlations are positive for both women and men under no feedback (ρ = .17 and .08 respectively, neither significant) as well as for men who receive feedback about their gender bias (ρ = .18, also insignificant) but negative for women who receive feedback (ρ = –.37), though this is also only suggestive (p = .23). We would cautiously interpret this result as meaning female subjects exposed to feedback try (but on average fail) to improve their gender bias between tests.

Fig. 2: Bias in response to feedback

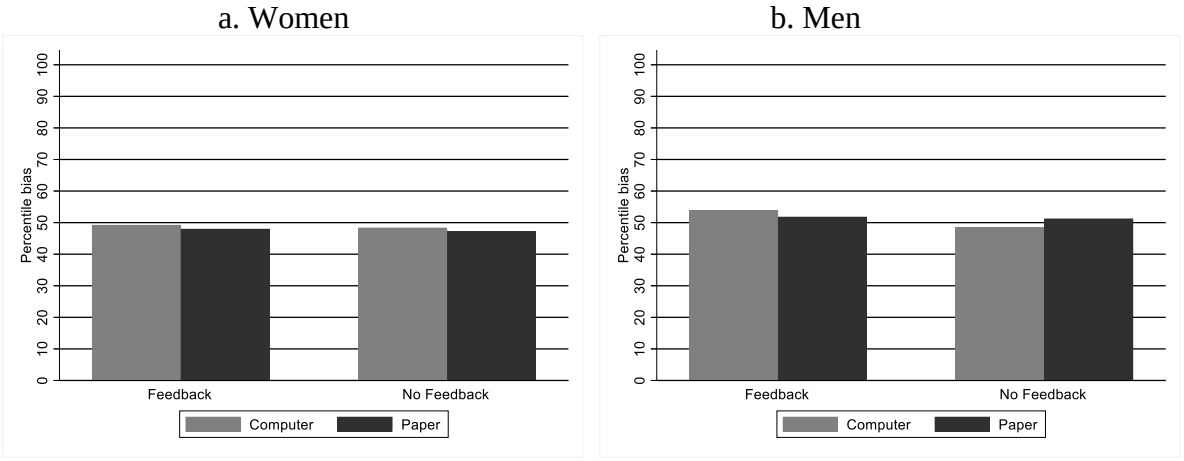
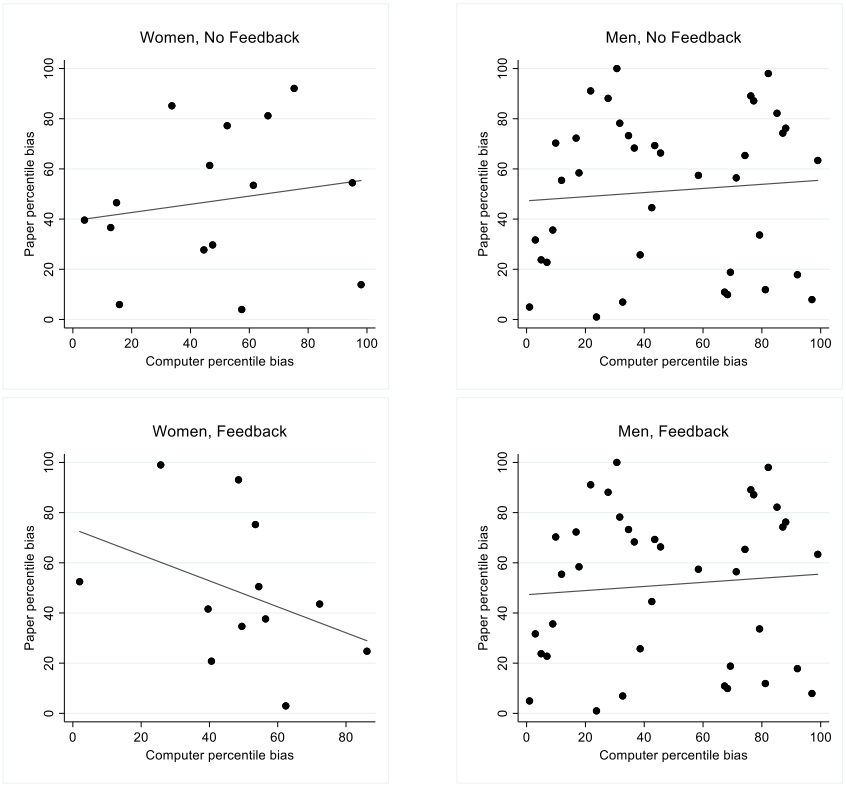
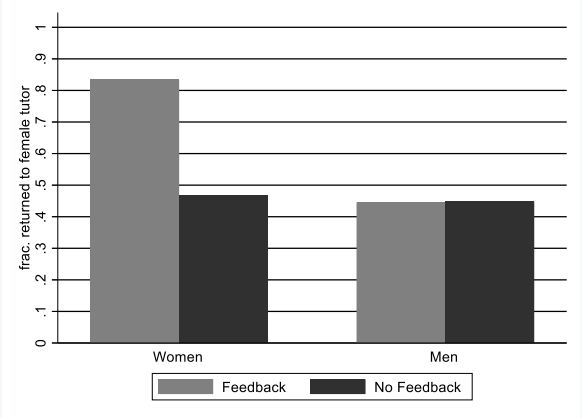


Fig. 3: Relationship between computerised and paper IAT percentiles for women and men by feedback condition



When we examine participants' likelihood to return their booklet to the female rather than the male tutor however, we see a stark pattern. This is indicated in Figure 4. Women receiving feedback were much more likely to return their IAT booklet to the female rather than the male tutor, compared with women who did not receive feedback on the computer IAT (ranksum p = .05). Men on the other hand were just as likely to return their booklets to the female tutor under both conditions (p = .98).

Fig. 4: Likelihood to return the paper IAT to female tutor



It seems from these aggregate patterns that among women at least, participants receiving feedback about their implicit gender bias made efforts to improve this when measured again; but ended up giving more work to the female tutor once they had demonstrated to themselves their improvement. The data in fact paint a more nuanced picture than this. Table 2, which reports the coefficients from linear probability regressions of returning the IAT booklet to the female tutor on feedback treatment, participant improvement score, and their interaction, broken out by participant gender, illustrates. We see among women the main effect of feedback for those whose bias percentile score did not improve is positive as per Figure 2. Note however the interaction between feedback and improvement. It is precisely those women whose bias percentile did *not* improve or indeed worsened who were most likely to hand their paper in to the female tutor. Male subjects on the other hand display no noticeable difference in behaviour between those whose bias scores improved or not, though these effects are very imprecisely estimated. It could be that since men on average did not improve with feedback, no moral licensing resulting from a "second chance" at the paper IAT would have been sought.

Tab. 2: Linear probability regressions of returning IAT to female tutor on feedback

	Women	Men
feedback	.373**	.008
	(.171)	(.117)
improvement	.005	-.004
	(.008)	(.004)
feedback ×	-.013	-.002
improvement	(.009)	(.006)
constant	0.467	0.442
	(.137)	(.082)

Concluding remarks

As ways of detecting the role of bias in everyday decisions and the cumulative effects it exerts on unequal outcomes continue to develop, we are beginning to understand better how to identify crucial points for policy intervention to contrast gender bias and stereotyping. The next stage for research is to conduct thorough assessments of the policies that work to counter unconscious bias, and their appropriateness for different policy areas. Our experiment in revealing bias, licensing bias (by eliciting it again) and exercising biased choices shows that revealing gender bias does not lead to corrective behaviour by men in our student sample, but it does on average lead to correction and thereafter to a larger gender biased choice by women in the sample. The experiment also suggests that the effects are different depending on the initial level of bias of the subjects as well as their gender. This kind of work needs to be replicated to much larger scales to be meaningful, but it does illustrate the need for caution when advocating bias revelation, as it might inadvertently cause people to act as if they have overcome their bias when they have not. Together with the gains that can be achieved by contrasting unconscious bias, work will anyway need to continue in combating structural barriers and institutional 'gendered bottlenecks' in order to realise progress towards realising gender equality**.**

References

Adamkovič, M. & Martončik, M. (2017). A review of consequences of poverty on economic decision-making: A hypothesized model of a cognitive mechanism. Frontiers in Psychology, 8, 1784.

Akerlof, G.A., 1980. A theory of social custom, of which unemployment may be one consequence. *The quarterly journal of economics*, *94*(4), pp.749-775.

Alan, S., Ertac, S., Kubilay, E. and Loranth, G., 2020. Understanding gender differences in leadership. *The Economic Journal*, *130*(626), pp.263-289.

Alan, S., Boneva, T. and Ertac, S., 2019. Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, *134*(3), pp.1121-1162.

Alan, S., Ertac, S. and Mumcu, I., 2018. Gender stereotypes in the classroom and effects on achievement. *Review of Economics and Statistics*, *100*(5), pp.876-890.

Alesina, A., Carlana, M., Ferrara, E.L. and Pinotti, P., 2018. *Revealing Stereotypes: Evidence from Immigrants in Schools*(No. w25333). National Bureau of Economic Research.

Anderson, L., Fryer, R. and Holt, C., 2006. Discrimination: experimental evidence from psychology and economics. *Handbook on the Economics of Discrimination*, pp.97-118.

Andreoni, J. and Vesterlund, L., 2001. Which is the fair sex? Gender differences in altruism. *The Quarterly Journal of Economics*, *116*(1), pp.293-312.

Apicella, Coren L., and Anna Dreber. 2015. "Sex Differences in Competitiveness: Hunter-Gatherer Women and Girls Compete Less in Gender-Neutral and Male-Centric Tasks." *Adaptive Human Behavior and Physiology* 1 (3): 247– 69.

Arrow, K. J. (1973), "The Theory of Discrimination", in O. Ashenfelter and A. Rees (eds.), *Discrimination in Labor Markets*, Princeton, NJ: Princeton University Press.

Ashenfelter, O. and Oaxaca, R., 1987. The economics of discrimination: Economists enter the courtroom. *The American Economic Review*, *77*(2), pp.321-325.

Atewologun, D., Cornish, T. and Tresh, F., 2018. Unconscious bias training: An assessment of the evidence for effectiveness. *Equality and Human Rights Commission Research Report Series*.

Bagues, M. & Esteve-Volart, B. (2010). Can gender parity break the glass ceiling? Evidence from a repeated randomized experiment. *Review of Economic Studies*, 77(4), 1301-1328.

Baldiga Coffman, K., 2014. Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, *129*(4), pp.1625-1660.

Becker, G.S., 1957. *The economics of discrimination*. University of Chicago press.

Bertrand, M. and Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American economic review*, *94*(4), pp.991-1013.

Bertrand, M., Goldin, C. and Katz, L.F., 2010. Dynamics of the gender gap for young professionals in the financial and corporate sectors. *American Economic Journal: Applied Economics*, *2*(3), pp.228-255.

Bian, L., Leslie, S.J. and Cimpian, A., 2017. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), pp.389-391.

Blumberg, R.L., 2015. Eliminating gender bias in textbooks: Pushing for policy reforms that promote gender equity in education. *Background paper for EFA Global Monitoring Report*.

Blank, A., Houkamau, C. and Kingi, H., 2016. *Unconscious bias and education: A comparative study of Māori and African American students*. Oranui Diversity Leadership.

Blanken, I., van de Ven, N., & Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*. 41(4), 540-558.

Bohnet, I. (2016). *What Works: gender equality by design.* Cambridge, MA: Belknap Press of Harvard University Press.

Bohren, J.A., Imas, A. and Rosenberg, M., 2019. The dynamics of discrimination: Theory and evidence. *American Economic Review*, *109*(10), pp.3395-3436.

Booth, A.L., 2009. Gender and competition. *Labour Economics*, *16*(6), pp.599-606.

Bordalo, P., Coffman, K., Gennaioli, N. and Shleifer, A., 2016. Stereotypes. *The Quarterly Journal of Economics*, *131*(4), pp.1753-1794.

Breda, T., Grenet, J., Monnet, M. and Van Effenterre, C., 2018. Can female role models reduce the gender gap in science? Evidence from classroom interventions in French high schools.

Brown, C.S. and Stone, E.A., 2016. Gender stereotypes and discrimination: how sexism impacts development. In *Advances in child development and behavior* (Vol. 50, pp. 105-133).

Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology*, 7, 765-786.

Burns, M.D., Monteith, M.J. and Parker, L.R., 2017. Training away bias: The differential effects of counterstereotype training and self-regulation on stereotype activation and application. *Journal of Experimental Social Psychology*, *73*, pp.97-110.

Campbell, T., 2015. Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment. *Journal of Social Policy*, *44*(3), pp.517-547.

Carlana, M. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *Quarterly Journal of Economics*, *134*(3), pp. 1163-1224.

Carrell, S.E., Page, M.E., West, J.E., 2010. Sex and science: how professor gender perpetuates the gender gap. Q. J. Econ. 125 (3), 1101–1144.

Carter, M.J., 2014. Gender socialization and identity theory. *Social Sciences*, *3*(2), pp.242-263.

Carvalho, L. S., Meier, S., & Wang, S. W. (2016). Poverty and economic decision-making: Evidence from changes in financial resources at payday. American Economic Review, 106(2), 260-284.

Cascio, Jessica and E. Ashby Plant Prospective moral licensing: Does anticipating doing good later allow you to be bad now? Journal of Experimental Social Psychology, Volume 56, 2015, pp. 110-116

Cialdini, R.B. and Trost, M.R., 1998. Social influence: Social norms, conformity and compliance.

Clot, S., Grolleau, G., & Ibanez, L. (2016). Do good deeds make bad people? *European Journal of Law and Economics*, 42(3), 491-513.

Cornwell, C., Mustard, D.B. and Van Parys, J., 2013. Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human resources*, *48*(1), pp.236-264.

Coffman, K.B., 2014. Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, *129*(4), pp.1625-1660.

Conway, P. & Peetz, J. (2012). When does feeling moral actually make you a better person? Conceptual abstraction moderates whether past moral deeds motivate consistency or compensatory behaviour. Personality and Social Psychology Bulletin, 38(7), 907-919.

Costa Jr, P., Terracciano, A. and McCrae, R.R., 2001. Gender differences in personality traits across cultures: robust and surprising findings. Journai of Personality and Social Psychology 81(2),pp.322-33.

Criado Perez, C.., 2019. *Invisible Women: Exposing data bias in a world designed for men*. Random House.

Dee, T. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95, 158-165.

Devine, P.G., Forscher, P.S., Austin, A.J. and Cox, W.T., 2012. Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology*, *48*(6), pp.1267-1278.

Dutt, K., Pfaff, D.L., Bernstein, A.F., Dillard, J.S. and Block, C.J., 2016. Gender differences in recommendation letters for postdoctoral fellowships in geoscience. *Nature Geoscience*, *9*(11), p.805.

Eccles, J.S., Freedman-Doan, C., Frome, P., Jacobs, J. and Yoon, K.S., 2000. Gender-role socialization in the family: A longitudinal approach. *The developmental social psychology of gender*, pp.333-360.

Eckel, C. C. and Fullbrunn, S. C., 2015. Thar SHE Blows? Gender, Competition, and Bubbles in Experimental Asset Markets. *American Economic Review*, 105(2):906–20.

Eswaran, M., 2014. *Why gender matters in economics*. Princeton University Press.

Equality Challenge Unit (2013). *Unconscious bias in Higher Education.*

Evans, G. W. & Schamberg, M. A. (2009). Childhood poverty, chronic stress, and adult working memory. Proceedings of the National Academy of Sciences of the United States of America, 106(16), 6545-6549.

Ewens, M., Tomlin, B. and Wang, L.C., 2014. Statistical discrimination or prejudice? A large sample field experiment. *Review of Economics and Statistics*, *96*(1), pp.119-134.

Falk, A., Becker, A., Dohmen, T.J., Enke, B. and Huffman, D., 2015. The nature and predictive power of preferences: Global evidence.

Favara, M., 2012. The Cost of Acting'Girly': Gender Stereotypes and Educational Choices.

Figlio, D. (2005). Names, expectation and the black-white test score gap. *NBER Working Paper*, 11195.

Fiske, S.T. and Stevens, L.E., 1993. *What's so special about sex? Gender stereotyping and discrimination*. Sage Publications, Inc.

FitzGerald, C., Martin, A., Berner, D. and Hurst, S., 2019. Interventions designed to reduce implicit prejudices and implicit stereotypes in real world contexts: a systematic review. *BMC psychology*, *7*(1), p.29.

Fryer Jr, R.G. and Levitt, S.D., 2010. An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), pp.210-40.

Gilliam, W.S., Maupin, A.N., Reyes, C.R., Accavitti, M. and Shic, F., 2016. Do early educators' implicit biases regarding sex and race relate to behavior expectations and recommendations of preschool expulsions and suspensions. *Research Study Brief. Yale University, Yale Child Study Center, New Haven, CT*.

Guiso, L., Monte, F., Sapienza, P. and Zingales, L., 2008. Culture, gender, and math. *Science*, *320*(5880), pp.1164-1165.

Glover, D., Pallais, A. and Pariente, W., 2017. Discrimination as a self-fulfilling prophecy: Evidence from French grocery stores. *The Quarterly Journal of Economics*, *132*(3), pp.1219-1260.

Gneezy, Uri, Muriel Niederle, and Aldo Rustichini. 2003. Performance in Competitive

Environments: Gender Differences. *Quarterly Jour- nal of Economics* 118 (3): 1049–74.

de San Román, A.G. and de La Rica, S., 2016. Gender gaps in PISA test scores: The impact of social norms and the mother's transmission of role attitudes. *Estudios de Economía Aplicada*, *34*(1), pp.79-108.

Groot, W. and van den Brink, H.M., 1996. Glass ceilings or dead ends: Job promotion of men and women compared. *Economics Letters*, *53*(2), pp.221-226.

Guryan, J. and Charles, K.K., 2013. Taste-based or statistical discrimination: The economics of discrimination returns to its roots. *The Economic Journal*, *123*(572), pp.F417-F432.

Greenwald, A. G., Bosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. Journal of Personality and Social Psychology, 85(2), 197-216.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, 4(4):146–68.

Holroyd, J. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition*, 33, 511-523.

Holroyd, J., Scaife, R., & Stafford, T. (2017). Responsibility for implicit bias. *Philosophy Compass*, 12(3), e12410.

Howell, J.L., Gaither, S.E. and Ratliff, K.A., 2015. Caught in the middle: Defensive responses to IAT feedback among whites, blacks, and biracial black/whites. *Social Psychological and Personality Science*, *6*(4), pp.373-381.

Jacob, B.A. and Wilder, T., 2010. *Educational expectations and attainment* (No. w15683). National Bureau of Economic Research.

Jassem. and K.D. Harber, (2005). 'Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies' *Personality and Social Psychology Review* 9 (2): 131–155.

Johns, M., Schmader, T. and Martens, A., 2005. Knowing is half the battle: Teaching stereotype threat as a means of improving women's math performance. *Psychological Science*, *16*(3), pp.175-179.

Jussim, Lee, Jarret Crawford, Stephanie Anglin, John Chambers, Sean Stevens, and Florette Cohen, ''Stereotype Accuracy: One of the Largest and Most Replicable Effects in All of Social Psychology,'' in The Handbook of Prejudice, Stereotyping, and Discrimination, Todd Nelson, ed. (Mahwah, NJ: Lawrence Erlbaum, 2015).

Kahneman, Daniel, and Amos Tversky, ''On the Psychology of Prediction,'' Psychological Review, 80 (1973), 237–251.

Kahneman, Daniel, 2011. Thinking, fast and slow. *Allen Lane, New York, USA*.

Karwowski, M., Lebuda, I., Wisniewska, E. and Gralewski, J., 2013. Big Five Personality Traits as the Predictors of Creative Self-Efficacy and Creative Personal Identity: Does Gender Matter?. *The Journal of Creative Behavior*, *47*(3), pp.215-232.

Kelly, D. & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass*, 3(3), 522-540.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Sunstein, C.R., 2019. *Discrimination in the Age of Algorithms* (No. w25548). National Bureau of Economic Research.

Knowles, J., Persico, N. and Todd, P., 2001. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, *109*(1), pp.203-229.

Koenig, A.M., 2018. Comparing prescriptive and descriptive gender stereotypes about children, adults, and the elderly. *Frontiers in psychology*, *9*, p.1086.

Krings, F., Johnston, C., Binggeli, S. and Maggiori, C., 2014. Selective incivility: Immigrant groups experience subtle workplace discrimination at different rates. *Cultural Diversity and Ethnic Minority Psychology*, *20*(4), p.491.

Kuhn, P., Villeval, MC. (2015). Are women more attracted to cooperation than men?, The Economic Journal, 125, 115–140. Doi: 10.1111/ecoj.12122.

La Ferrara, Eliana. "Aspirations, Social Norms, and Development." *Journal of the European Economic Association* 17, no. 6 (2019): 1687-1722.

Lewis, G.B., 1986. Gender and promotions: Promotion chances of white men and women in federal white-collar employment. *Journal of Human Resources*, pp.406-419.

Lemm, K. M., Lane, K. A., Sattler, D. N., Khan, S. R., & Nosek, B. A. (2008). Assessing implicit cognitions with a paper-format Implicit Association Test. In: *The Psychology of Modern Prejudice*, M. A. Morrison & T. G. Morrison (Eds.), pp. 123-146. Hauppauge, NY: Nova Science Publishers.

Lichand, G. & Mani, A. (2020). Cognitive droughts. University of Zürich Working Paper No. 341.

Lieberman, M.D., 2013. *Social: Why our brains are wired to connect*. OUP Oxford.

List, J.A., 2004. The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics*, *119*(1), pp.49-89.

Lowes, S., Nunn, N., Robinson, J. A., and Weigel, J. (2015). Understanding ethnic identity in Africa: Evidence from the Implicit Association Test (IAT). *The American Economic Review*, 105(5):340–45.

Madva, A. (2017). Implicit bias, moods, and moral responsibility. *Pacific Philosophical Quarterly*, 99(S1), 53-78.

Magua, W., Zhu, X., Bhattacharya, A., Filut, A., Potvien, A., Leatherberry, R., Lee, Y.G., Jens, M., Malikireddy, D., Carnes, M. and Kaatz, A., 2017. Are female applicants disadvantaged in National Institutes of Health peer review? Combining algorithmic text mining and qualitative methods to detect evaluative differences in R01 reviewers' critiques. *Journal of Women's Health*, *26*(5), pp.560-570.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. Science, 341(6149), 976-980.

Marx, D.M. and Roman, J.S., 2002. Female role models: Protecting women's math test performance. *Personality and Social Psychology Bulletin*, *28*(9), pp.1183-1193.

Mazar, N. and Zhong, C.B., 2010. Do green products make us better people?. *Psychological science*, *21*(4), pp.494-498.

Merritt, Anna, Daniel Effron, Steven Fein, Ken Savitsky, Daniel Tuller, and Benoît Monin (2012) The strategic pursuit of moral credentials *Journal of Experimental Social Psychology,* Volume 48, Issue 3, 2012, pp. 774-777

Monin, B. & Miller, D. T., 2001. Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology*, *81*(1), pp.33-43.

Morris, E.W. and Perry, B.L., 2017. Girls behaving badly? Race, gender, and subjective evaluation in the discipline of African American girls. *Sociology of education*, *90*(2), pp.127-148.

Mullen, E. & Monin, B. (2016). Consistency versus licensing effects of past moral behaviour. *Annual Review of Psychology*, 67, 363-385.

Nelson, J.A., 2014. The power of stereotyping and confirmation bias to overwhelm accurate assessment: the case of economics, gender, and risk aversion. *Journal of Economic Methodology*, *21*(3), pp.211-231.

Nelson, J.A., 2015. Are Women Really More Risk-Averse than Men? A Re-Analysis of the Literature Using Expanded Methods. *Journal of Economic Surveys*, *29*(3), pp.566-585.
Nelson, T.D., 2009. *Handbook of prejudice, stereotyping, and discrimination*. Psychology Press.

Oaxaca, R., 1973. Male-female wage differentials in urban labor markets. *International economic review*, pp.693-709.

Oxoby, R.J., 2014. Social inference and occupational choice: Type-based beliefs in a Bayesian model of class formation. *Journal of Behavioral and Experimental Economics*, *51*, pp.30-37.

Phelps, Edmund (1972). "The Statistical Theory of Racism and Sexism". *The American Economic Review*. **62** (4): 659–661.

Piirto, J., 1991. Why are there so few?(Creative women: Visual artists, mathematicians, musicians). *Roeper Review*, *13*(3), pp.142-147.

Porter, C. and Serra, D., 2019. Gender differences in the choice of major: The importance of female role models. *American Economic Journal: Applied Economics*.

Prentice, D.A. and Carranza, E., 2002. What women and men should be, shouldn't be, are allowed to be, and don't have to be: The contents of prescriptive gender stereotypes. *Psychology of women quarterly*, *26*(4), pp.269-281.

Qian, M.K., Quinn, P.C., Heyman, G.D., Pascalis, O., Fu, G. and Lee, K., 2017. Perceptual individuation training (but not mere exposure) reduces implicit racial bias in preschool children. *Developmental psychology*, *53*(5), p.845.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales. 2015. "Taste for Competition and the Gender Gap Among Young Business Pro- fessionals." National Bureau of Economic Research Working Paper 21695.

Rienzo, C., Rolfe, H. and Wilkinson, D., 2015. Changing Mindsets: Evaluation Report and Executive Summary. *Education Endowment Foundation*.

Rippon, G., 2019. *The Gendered Brain: The new neuroscience that shatters the myth of the female brain*. Random House.

Rodgers, W.M., 2009. *Handbook on the Economics of Discrimination*. Edward Elgar Publishing.

Sanchez Perry, Holly (2015) Ensuring the Success of Early Childhood Education: Confronting Implicit Biases Comparative Education Law and Policy: Early Childhood Education.

Schlesinger, A., O'Hara, K.P. and Taylor, A.S., 2018, April. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 315). ACM.

Schmitt, D.P., Realo, A., Voracek, M. and Allik, J., 2008. Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of personality and social psychology*, *94*(1), p.168.

Schmader, T., 2010. Stereotype threat deconstructed. *Current Directions in Psychological Science*, *19*(1), pp.14-18.

Schneider, David, The Psychology of Stereotyping (New York: Guilford Press, 2004).

Shah, A. K., Mullainathan, S., & Shafir, E. (2012). Some consequences of having too little. Science, 338(6107), 682-685.

Shenouda, C.K. and Danovitch, J.H., 2014. Effects of gender stereotypes and stereotype threat on children's performance on a spatial task. *Revue internationale de psychologie sociale*, *27*(3), pp.53-77.

Simbrunner, P. & Schlegelmilch, B. B. (2017). Moral licensing: A culture-moderated meta-analysis. *Management Review Quarterly*, 67, 201-225.

Sprietsma, M. (2009). Discrimination in grading? Experimental evidence from primary school. ZEW Discussion Paper, 09-074.

Sullivan, O., Gershuny, J. and Robinson, J.P., 2018. Stalled or Uneven Gender Revolution? A Long-Term Processual Framework for Understanding Why Change Is Slow. *Journal of Family Theory & Review*, *10*(1), pp.263-279.

Stiglitz, J.E., 1973. Approaches to the Economics of Discrimination. *The American Economic Review*, *63*(2), pp.287-295.

Teal, C.R., A.C. Gill, A.R.Green and S. Crandall (2012). Helping Medical Learners Recognise and Manage Unconscious Bias Toward Certain Patient Groups *Medical Education* 46 (1): 80-88.

Tversky, Amos, and Daniel Kahneman, ''Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment,'' Psychological Review, 90 (1983), 293–315.

Van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., and Holland, R. W. (2010). The implicit prejudiced attitudes of teachers: Relations to teacher expectations and the ethnic achievement gap. *American Educational Research Journal*, 47(2):497–527.

Whiteout, S. (2018). Popularizing Wokeness. *Harvard Journal of African American Public Policy*, 2017-18, 63-70.

Williams, W.M. and Ceci, S.J., 2015. National hiring experiments reveal 2: 1 faculty preference for women on STEM tenure track. *Proceedings of the National Academy of Sciences*, *112*(17), pp.5360-5365.

Wood, M., Hales, J., Purdon, S., Sejersen, T. and Hayllar, O., 2009. A test for racial discrimination in recruitment practice in British cities. *Department for Work and Pensions Research Report*, *607*.

Wu, A.H., 2018. Gendered Language on the Economics Job Market Rumors Forum. In *AEA Papers and Proceedings* (Vol. 108, pp. 175-79).

Zetland, D. and Della Giusta, M., 2011. *Focal points, gender norms and reciprocation in public good games* (No. em-dp2011-01). Henley Business School, Reading University.