Chowdhury, Shyamal; Klauzner, Ilya; Slonim, Robert

Working Paper

What's in a Name? Does Racial or Gender Discrimination in Marking Exist?

IZA Discussion Papers, No. 13890

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Chowdhury, Shyamal; Klauzner, Ilya; Slonim, Robert (2020) : What's in a Name? Does Racial or Gender Discrimination in Marking Exist?, IZA Discussion Papers, No. 13890, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/232642

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13890

What's in a Name? Does Racial or Gender Discrimination in Marking Exist?

Shyamal Chowdhury
Ilya Klauzner
Robert Slonim

NOVEMBER 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13890

What's in a Name? Does Racial or Gender Discrimination in Marking Exist?

Shyamal Chowdhury
University of Sydney and IZA

Ilya Klauzner
BOCSAR

Robert Slonim
University of Sydney and IZA

NOVEMBER 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

What's in a Name? Does Racial or Gender Discrimination in Marking Exist?*

We study whether racial or gender discrimination in marking exists at universities by conducting an experiment at a major Australian university where we randomly assigned names indicative of White, Chinese or Adopter identities (comprised of a White first name and Chinese surname) and male or female gender to real exam coversheets and recruited university graders to mark these exams. We find that the most economically-significant evidence of discrimination is found at grade thresholds. Exam scripts with Chinese and Adopter names are less likely than White names to receive a mark just above a grade threshold. Conversely, scripts with Chinese names receive a small marking bonus on average compared to the same script with a White name. Discrimination at grade thresholds is found to be more consistent with taste-based discrimination, whereas discrimination at the average is more consistent with statistical discrimination.

JEL Classification: J15, J71, J64, C93

Keywords: racial discrimination, experiment, marking

Corresponding author:
Shyamal Chowdhury
University of Sydney
NSW 2006
Australia

E-mail: shyamal.chowdhury@sydney.edu.au

* The analysis has benefited from many comments on presentations of the findings at the University of Sydney. We thank Medha Sengupta for her excellent editorial assistance, and Alex Berger, Sam Batchelor, Liliana Tai, and Patrick Hendy for their help with conducting the experiment. We also acknowledge the academics who taught the courses in our experiment, and who gave up their time to make this experiment possible. Their names are left out for privacy concerns. The study received ethics approval from the University of Sydney (project 2018/792).

1 Introduction

Pay discrimination against minorities and female workers is widely observed in many societies.[1] One plausible explanation for at least part of the pay gaps is the racial and gender discrimination faced by minority and female workers in the labour market. (Bertrand & Mullainathan, 2004; Goldin & Rouse, 2000). However, discrimination against minorities can extend far beyond the labour market. Discrimination permeates through many aspects of society, from the criminal justice system [2] to who gets let on a bus for free.[3]

Discrimination can also be present within the educational system. Harvard University, for example, has been embroiled in a law suit alleging that its admissions policy is discriminatory towards Asian Americans (Hartocollis, 2019). Further, Harvard and the University College London have adopted strategies for 'inclusion and belonging' and 'Race Equality' respectively, suggesting that they are aware of the existence of discrimination within their institutions and are taking active steps to address it. (Presidential Task Force on Inclusion and Belonging, 2018; *Race Equality Charter*, 2020).

Analogous to the racial and gender wage gaps that have been extensively studied as forms of discrimination in the labour market, disparities in academic achievement along racial lines are also seen in higher education. For example, university administrative data provided to us shows that students with Chinese names get notably worse marks on average than students with White names. In two of the three courses for which we received administrative data, students with Chinese names received on average 14.1 and 6.6 marks less than students with White names, which is 0.6 and 0.4 of a standard deviation, respectively.[4] Mark inequalities could exist for a variety of reasons but the existence of such a large gap highlights the need to determine whether racial discrimination is an underlying factor.

Discrimination based on observable characteristics which are not associated with productivity is a major issue in the labour market (e.g., Neal and Johnson 1996; Altonji and Blank 1999; Lang and Lehmann 2012; Neumark, Burn and Button 2019).[5]

[1] In the US, for example, Black men earn 20 percent less than White men and the median earnings of Hispanics are 32 percent less than that of Whites. Moreover, the gap between male and female earnings is still around 20 percent (Neumark, 2018).

[2] For example, high-profile policies like 'stop-and-frisk' have been ruled discriminatory by a judge (*Floyd, et al. v. City of New York, et al*, 2013) because half of all those searched were Black, and a third were Latino.

[3] Mujcic and Frijters (2013) found that Black and Indian passengers were much less likely to get a free bus ride in Brisbane, Australia

[4] However, given our administrative data was restricted to certain subjects, we do not know whether these gaps exist across the whole university

[5] See Bertrand and Duflo (2017) for a recent review of correspondence and audit studies in the context of labour markets.

In this paper, we examine whether racial minorities and females are discriminated against when being awarded marks at university. This is an important context to focus on given the university environment functions as the precursor to the labour market. Unlike studies on discrimination in the labour market, however, studies on discrimination in education (especially in regards to disparities in grading) are much less common.[6] However, if discrimination exists in university marking, minority candidates are disadvantaged from the start, which can go on to have significant adverse impacts on their entry into the labor market, as well as their lifetime earnings. Moreover, discrimination has a negative effect on the mental and physical health of discriminated groups and is associated with higher stress levels and health disparities (American Psychological Association, 2016).

From a purely economic point of view, discriminatory university marking clearly leads to inefficient outcomes for firms, as the signal that employers receive from a job applicant's academic achievements becomes noisy. This is because the best or most productive students don't receive marks that are commensurate with their level of academic achievement. From the perspective of employees, discrimination in marking means fewer economic opportunities and reduced social mobility for minority groups who often see education as their sole pathway out of poverty. Moreover, even in the absence of labour market discrimination, the presence of discrimination in educational settings such as universities would mean that a rational profit-maximizing employer could make incorrect inferences about the educational performance of minority candidates', as their educational record would reflect discriminatory marking rather than their actual academic performance.

The study of discrimination has a strong theoretical basis in Economics. Becker's (1957) theory of taste-based discrimination analyses discrimination in the labour market in terms of preferences, with some employers getting dis-utility from interacting with employees from minority groups. As a result, they may not hire them, or if they do, offer them lower wages than they would offer to employees from the majority group for the same level of productivity. Alternatively, Phelp's (1972) and Arrow's (1972) theories of statistical discrimination view discrimination as a result of rational profit-maximising behaviour in a world of imperfect information. Lacking full information about the productivity of an employee, employers supplement this with information about their average group productivity. If minority applicants are less productive on average, employers may favor majority applicants as their expected productivity is higher.

Finally, implicit discrimination is a psychological concept which theorises discrimination as an unconscious process, in contrast to the explicit processes associated with economic theories. Implicit attitudes are particularly likely to arise for tasks that involve high cognitive loads, time pressure, and ambiguity,

[6]Exceptions include Lavy (2008); Breda and Ly (2015); Feld, Salamanca, and Hamermesh (2016); Hanna and Linden (2012); Sprietsma (2013); Van Ewijk (2011).

such as marking. This means that markers are more likely to rely on their 'heuristics' or stereotypes when assessing academic work. (Bertrand, Chugh, & Mullainathan, 2005).

The aforementioned theoretical models of discrimination help to make testable predictions in the context of marking. Taste-based discrimination implies animus to minorities, and consequently, bias against minority students (assuming that the majority of markers are not members of minority groups). However, statistical discrimination, wherein markers take into account a group's average mark, could benefit minority students and women in subjects where they are known to perform better on average. Finally, implicit bias on the part of markers would mean markers rely on their 'heuristics' and common stereotypes about the academic performance of minority student and females in different subjects.

To test the above predictions, we conducted a randomized control trial (RCT) at a major Australian university. Real exam scripts were photocopied and names indicative of different ethnicities and genders were randomly added to their coversheets and given to different markers to grade. The names could be categorised as White, Chinese or Adopter,[7] and male or female. Markers were not aware that they were participating in a research study during the course of the experiment, and were instead told that they were marking exams for credit. Given that the names were randomly assigned to exam coversheets and that all markers graded the same exam papers but with different names on the coversheet, any differences in marks awarded can be attributed to racial and/or gender discrimination.

The literature on marking bias is relatively nascent.[8] Using a difference-in-difference design, Lavy (2008) tested the existence of gender-based marking bias in Israel. He compared a non-blind score that a student received from their teacher with a blind score that s/he received on a state exam. The blind score was used as a counterfactual as it should not have been susceptible to bias. He found evidence for a bias against boys which was as high as 0.272 of a s.d. Breda and Ly (2015) used a similar strategy, comparing the blind written score with a non-blind oral score on an entrance exam to an elite French university. They found that females were favoured by examiners in more male-dominated fields, whereas males were favoured in more female-dominated fields. Switching from biology (a field considered more feminine) to math (a field considered less feminine) led female candidates to gain 0.25 of a s.d. in oral tests compared to written tests.

[7] Names comprised of a White first name and a Chinese last name. See Chowdhury, Ooi, and Slonim (2020) for details.

[8] Studies examining biases in exam marking have an important observational advantage over studies looking at discrimination in employer responses to job applications. In job applications, correspondence studies (e.g., Bertrand and Mullainathan (2004)) can only go as far as observing who gets an interview, and thus one cannot directly infer the final outcome: whether there is bias in getting a job. However, the final outcome of marking can be easily observed as it is simply the grade each student receives.

However, the difference-in-difference technique that the aforementioned papers rely on to identify discrimination has some important limitations. In Lavy (2008), for example, whereas the blind score was marked externally, the non-blind score was marked by the student's teacher. This meant that any observed bias (as indicated by discrepancies between the two scores) could not be easily categorised as being based on gender and/or race, as a teacher could favour or disfavour a particular student based on other factors (e.g. the student's personality, behaviour, etc.) Secondly, in Lavy (2008) and Breda and Ly (2015), the blind assessments differed from the non-blind assessments, meaning that the form of the assessments and the skills they measured might have affected girls and boys differently. For example, blind external exams may be more high-stakes than internal exams, and so boys may do better in these exams because they tend to be more competitive.[9] Thirdly, the notion of the 'stereotype threat' could have played a role in the aforementioned studies. If a stereotype about the academic performance of a certain group already exists, members of the group can change their behaviour and perform negatively in response to the stereotype. However, this behaviour is only expected to be observed in non-blind exams. Our study overcomes these limitations as all markers in each subject marked the exact same exam papers, and all students took the exam believing their paper would be marked blindly.

To the best of our knowledge, our paper is the only experimental study to look at marking bias in a university context.[10] In India, Hanna and Linden (2012), created and implemented an exam competition for children between the ages of 7 and 14, and then hired markers to grade the exams which had coversheets with randomised child characteristics including age, gender and caste. They found that the marks awarded to children from 'lower castes' were about 0.03 to 0.08 s.d. lower than those given to students from 'higher castes'. Similarly, Sprietsma (2013) randomly assigned German-sounding or Turkish-sounding names to primary school essays, and had them marked by different teachers in Germany. She used a two-way fixed effects strategy of teacher and essay fixed effects which is also employed in our study. She found that essays assigned Turkish-sounding names received a grade that was 10% of a s.d. lower than essays assigned German-sounding names, a small but significant decrease. Conversely, Van Ewijk (2011) who conducted a similar experiment but with names indicative of Dutch, Turkish or Moroccan backgrounds, found no differences in the grades awarded to the three experimental groups.

Our identification strategy improves upon previous experimental studies. Markers were not told that they were participating in an experiment until after the experiment ended, removing any concerns about markers being vulnerable to social desirability bias, and simply marking how they believed experimenters wanted them to mark. In contrast, the markers in both Sprietsma (2013) and Van Ewijk (2011) knew

[9] For a summary of competitiveness and the gender gap in achievement see Niederle and Vesterlund (2010)

[10] Feld et al. (2016) is one exception, but dealt with a slightly different research question, trying to distinguish between those that favour their own kind and those that discriminate against others.

they were participating in an experiment (albeit were not privy to the aim), which could have exposed the studies to social desirability bias. Additionally, we collected data about markers including the number of semesters they had previously worked as a marker, and their ethnicity, in order to to determine whether the results were consistent with taste-based or statistical discrimination.

Further, examining discrimination in university marking can provide important insights into discrimination in the labour market, and Australian universities provide an interesting context in which to examine this discrimination. Australia is a diverse and multicultural country, and Australian universities have a large international student population, with students from China making up 10% of all students. How discrimination manifests in a multicultural context with a significant population of Chinese international students is an interesting question. Chowdhury et al. (2020) found that Chinese-Australian university graduates are employed less and have significantly lower earnings than White-Australians, even though Chinese-Australians are over-represented at universities. Moreover, university administrative data examined here shows that students with Chinese names get worse marks on average than White students. Therefore, examining whether discrimination is already occurring in university settings is a pertinent question. Finally, following Chowdhury et al. (2020), we also examine whether the extent of discrimination can be attenuated if the student 'adopts' a White first name. Chowdhury et al. (2020) found that having an 'Adopter' name dramatically reduced the level of discrimination against Chinese job applicants. Adopting a White name may signal integration, and Chinese-Australians who choose to do so may consequently face less taste-based discrimination than those who retain both Chinese first and last names.

We find that discrimination in university marking occurs 'where it matters'. The most economically-significant evidence of discrimination occurs at the pass-fail threshold where Chinese and Adopter names are, up to 26.1 and 19.5 percentage points less likely to receive a mark just above a grade threshold. However, this effect disappears when looking at differences in marks at the average.

The rest of this paper is organised as follows: Section 2 describes the experimental design, while Section 3 describes the identification strategy and reports the empirical results. Finally, Section 4, explores the results further and examines an administrative dataset for additional information. Section 5 concludes.

2 Experimental Design

We first selected three large junior units of study in the subject areas of Economics, Business and Information Technology (IT). We then randomly selected 101 real exam scripts, photocopied them, randomly

assigned names to their coversheets, and had them marked by 14 different markers: five markers each for Economics and Business, and four for IT.[11] Markers were told that they were marking replacement or external examinations. Each marked all the exam scripts for one subject. Therefore, each marker received an identical bundle of exam scripts for their particular subject, which differed only by the name (or lack thereof) on the coversheet. Table 1 shows the distribution of markers per condition.

[Table 1 here]

The exams included in our experiment accounted for a sizeable proportion of a student's final mark so were considered 'high-stakes'. The Economics exam was worth 25% of a student's final mark, while the Business and IT exams accounted for 55% and 60%, respectively. [12] 30 distinct exam scripts were randomly selected from Economics, while 36 and 35 exams were selected from Business and IT respectively. The random selection of these exams was stratified by exam quality as determined by past marks. Students from each subject were ordered by their cumulative score in the course at that point and divided into thirds. One third of the exam scripts were then randomly selected from each third. Business and Economics exam papers were each photocopied five times, whereas each IT paper was photocopied four times (one for each marker), similar to Hanna and Linden (2012).

2.1 Assignment of Ethnicity and Gender

The coversheets of the original exam scripts were removed and identical coversheets containing ethnically meaningful names of fictitious students were added. The names were randomly chosen from a name bank identical to the name bank in Chowdhury et al. (2020), containing 14 common male first names, female first names, and White and Chinese last names.[13] There were 6 possible treatment groups varying by gender and ethnicity: White female; White male; Chinese female; Chinese male; Adopter female; and Adopter male. As noted earlier, an Adopter name is one that consists of a White first name and a Chinese surname

[11] Additionally, four graders marked unnamed exams. More information is provided in Appendix Section A.5

[12] Whereas the exams for both Economics and IT were graded in full, only the short and extended response questions from the Business paper were marked (constituting 55% of the entire paper.

[13] White first names were selected from the most popular baby names in the birth registries of the New South Wales and Victorian state governments from the decade 1990-2000. White last names were taken from the 2007 government list of most common last names in Australia. Chinese first names were randomly selected from the website 'Top 100 Baby Names' and Chinese last names were based on household data collected from the Ministry of Public Security (China). Example names from each group are given as follows: Matthew Smith (White male); Sarah Jones (White female); Ming Li (Chinese male); Ying Wang (Chinese female); James Zhang (Adopter male), and Jessica Liu (Adopter female).

(e.g. James Zhang). Each marker received all 6 treatment groups in their bundle, which were randomly selected without replacement from the name bank to ensure a marker never received two identical names.

As much as possible, each marker graded an equal number of exam papers from each treatment group, stratified by exam quality e.g. if a marker received 36 exams, s/he would have received an equal number of treatment groups from the bottom, middle and top thirds of the exams, meaning they would have received 2 exams for one treatment group from each third. Each distinct exam script was assigned to as many treatment groups as possible, and the resulting papers were then given to different markers. For example, given there were five markers each for both Economics and Business, five names from five different treatment groups could be given to the same exam script and then assigned to different markers. One treatment name had to be excluded from each distinct exam script, which was done in an alternating order. As there were four markers for IT, two treatment names had to be excluded from each distinct exam script, which was also done in an alternating order. Therefore, every marker in each subject marked the same exams but with different names on the coversheets, and every distinct exam script had multiple names from different name groups attributed to it.

The exact coversheets were made to look as similar as possible to the official coversheets used for replacement exams for the three subjects. An example of a typical coversheet used in the experiment is shown in Figure 1. The name featured prominently in all coversheets.

2.2 Recruitment of Markers

Markers were selected according to the standard process for selecting markers in the schools of Economics, Business and IT. As is standard practice, the relevant course coordinator or teaching coordinator selected the markers. They sent an email to a pool of potential markers from each school, asking for expressions of interest. The potential pool of markers normally consists of casual academics or tutors who teach subjects within the school. The only restriction we placed on the selection of markers was that they could not be the official markers for the course, e.g. if a marker was scheduled to mark the Business final exam, then s/he was excluded from marking the 36 exams in our experiment. Given our experimental exam papers were taken from the actual exam papers for the three subjects, this restriction ensured that markers who had previously graded the original paper were not included in our experiment.

Markers were not told that they were taking part in an experiment, and instead believed they were marking for credit.[14] Markers for the Business and Economics exams were told they were marking re-

[14] As part of an ethics approval process, markers were only informed of their participation in the experiment following the

placement exams whereas those marking the IT papers were told they were marking external exams.[15] Following Hanna and Linden (2012), we wanted markers to believe they would have a direct impact on students' welfare, so we ensured the marks for each exam constituted a significant proportion of the final mark for the corresponding subject. Further, as markers did not know that they were taking part in an experiment, there was no risk of them displaying social desirability bias by changing their behaviour because they were being observed. Therefore, our design is much more robust than Van Ewijk (2011), for example, where markers knew they were participating in an experiment and were not marking for credit.

Markers received all communication about the marking process from the course coordinators,[16] and were given packets of exams that all had names on them.[17] This is a limitation of Feld et al. (2016) where markers could receive bundles that had both non-blind and blind exams which could have made markers suspicious. To replicate marking conditions as realistically as possible, course coordinators were instructed to give markers a deadline similar to what final exam markers typically receive. This tended to be about a week.

2.3 Summary Statistics and Internal Validity

Our random assignment ensured that the six treatment groups (White male, White female, Chinese male, Chinese female, Adopter male and Adopter female) were almost equally distributed across subjects and markers, as shown in columns 1-3 in Table 2 (and 3). We also collected administrative data to examine if the selection of our treatments aligned with the actual gender and ethnic distribution in the student population. This is presented under 'Official Names in columns 4-6 , and under 'Preferred Names' (indicating the students' nominated names) in columns 7-9. When comparing our experimental assignments

completion of marking and the collection of results.They were all given the opportunity to opt out and have their results removed from the analysis. However, no markers decided to opt out of our study.

[15] At no point during or after the experiment did we receive any feedback from any of the markers that suggested they thought they were not marking real exams.

[16] One course coordinator reported that two markers had begun to grade the exams together. We examine whether this would have affected the results in Section C.1, Table C4.

[17] It is important to note here that the university this experiment was conducted in switched to a policy of anonymous marking in 2018. Thus, the fact that the exam scripts contained names might have alerted the markers about the experiment, and thus constituted a flaw in our experimental design. However, given that all three course coordinators stated that they had previously marked replacement and external exams by themselves, the markers in our experiment had no prior experience with marking these types of exams, and thus no prior beliefs about whether these exams should have names on them. Consequently, they would have been less suspicious of the aspects specific to this experiment that distinguished it from the standard marking process, such as the fact that they were marking a limited quantity of papers which were photocopied and contained students' names.

to the distribution of official names in the administrative data, we see that White females and Adopters are over-represented in our experiment, whereas Chinese females tend to be under-represented. Moreover, in comparison to the distribution of preferred names in the administrative data Chinese names in our experimental groups are over-represented while Adopter names are no longer over-represented. However, our six treatment groups are well-aligned to the actual distributions of the student population of the three subjects overall. Finally, summary statistics of certain characteristics of the markers, such as the number of semesters that they had spent marking prior to the experiment and their ethnicity, are shown in Table 3.

[Table 2 here]

[Table 3 here]

2.4 Power

Lastly, we calculated the statistical power of our experiment to ascertain the minimum detectable effect size of our experiment. We used power simulations to calculate power across the two dimensions of fixed effects, exam scripts and markers. To do this, we simplified our design and looked at the ability of our experiment to detect an effect size between just two treatment groups, and we simulated two random shocks. The first was a random shock given to each marker, and the second was a random shock given within each exam script, varying by the name group the exam script was given. The full details of how the power simulation was completed can be found in the Appendix Section A.2. Our power calculations varied both the number of markers and the number of exam scripts. They show that our experiment is able to detect an effect size of 1.5 marks out of 100 when looking at all subjects, or 2 marks out of 100 when looking at any individual subject. Figures showing the power calculations are shown in Appendix Section A.2.

3 Empirical Estimation and Results

3.1 Identification Strategy

In order to identify the existence of racial and gender discrimination in marking, the primary empirical specification that we use is a two-way fixed effects model that takes the following form:

$$Y_{sm} = \beta_0 + \beta_1 chinese_s + \beta_2 adopter_s + \beta_3 female_s + \alpha_s + \alpha_m + \epsilon_{sm} \quad (1)$$

Where Y_{sm} is the dependent variable, which is either $Mark_{sm}$, the mark assigned to exam script s by marker m, or $JustAbove_{sm}$, a dummy variable equal to one if a script s received a mark just above a grade threshold. $chinese_s$ is equal to one for an exam script with a Chinese first name and Chinese last name, and $adopter_s$ is equal to one for an exam script with a White first name and Chinese last name. $female_s$ is equal to one for an exam script with a female first name. α_s refers to exam script fixed effects, and α_m refers to marker fixed effects. As the same script was marked multiple times, and the same marker graded many different name groups, we control for unobservables in exam quality and marker leniency. The comparison group are White Males. Here, β_1 and β_2 give estimates of discrimination towards Chinese and Adopter names respectively, compared to White names, and β_3 gives an estimate of discrimination against female names, compared to male names. As the assignment of names was randomised, and much of the unobservables controlled through script and marker fixed effects, the calculated coefficients can point towards a causal relationship.

3.2 Average Effects

We start with estimation equation 1, where we regress the mark an exam script received out of 100[18] on race and gender. Table 4 reports the results. The main coefficients of interest are 'Chinese', 'Adopter', and 'female', which are dummy variables equal to one if the exam script had a name indicative of the respective ethnicity or gender. The omitted group for all specifications is White males. Column (1) shows the raw differences in marks among the different groups, and excludes any marker characteristics or fixed effects. Column (2) adds controls of certain marker characteristics, the order the exam was returned in[19], and subject fixed effects. Column (3) adds exam script fixed effects, which controls for the unobservable differences in the quality of each exam script. Column (4) has both exam script and marker fixed effects to control for unobservable differences in the marks that different markers give.

[Table 4 here]

The results from the first three specifications are similar. There are no statistically significant differences

[18] If an exam was not marked out of 100 points, we convert its score to a score out of 100.

[19] Controls include the ethnicity and gender of the marker, the number of semesters they have been employed as markers, and the order in which they returned their exam scripts. Hanna and Linden (2012) show that all these variables have the potential to influence the grading decision.

in marks for any given ethnicity or gender compared to that of White males, suggesting that direct bias in the form of an explicit mark difference doesn't affect students with Chinese, Adopter, or female names. However, in model (4), the preferred model with both exam script and marker fixed effects, the coefficient for the Chinese group is approaching statistical significance with a p-value of 0.105. This implies a small positive bias of 1.14 marks (or 0.06 of a standard deviation) towards Chinese names. In the context of our power calculations, this effect size is just below our minimum detectable level, so a p-value of 0.105 may be considered low in this context. This is further discussed in Section 4.3.1. The results of column (4) are shown graphically in Figure 2.

[Figure 2 here]

The signs of the coefficients of each of the name categories are the same for the four models. If there was any difference in marks, we would expect exams with Chinese and Adopter names to have slightly higher marks compared to that of White males, whereas exams with female names would be awarded slightly lower marks. Further, the markers do not seem to distinguish between exams with Chinese and Adopter names. The 'Test of Adopter=Chinese' row shows the p-value from a test of equality of the coefficients of the Adopter and Chinese groups. The smallest p-value is 0.19, meaning we cannot reject the null that these coefficients are equal.

[Table 5 here]

We now turn to each subject and estimate equation 1 separately for each subject. Results are reported in Table 5. Column (1), Column (2) and Column (3) restrict the sample to Economics, Business and IT, respectively. All specifications have exam script and marker fixed effects. The coefficients for each group in Business and Economics has the same sign, while IT has negative signs for the Chinese and Adopter groups. The coefficient for the Chinese group is positive and statistically significant in Business. This suggests the existence of positive bias in the marking of this subject, with scripts with Chinese names receiving 2.24 extra marks, compared to the exact same script with a White name. This is equivalent to 0.11 of a standard deviation. However, even in this case, the p-value from a test of equality of coefficients between the Chinese and Adopter groups is 0.15, meaning we are unable to reject the null that these coefficients are the same. Tables 4 and 5 both have standard errors that are not clustered, as assignment to treatment is not clustered (i.e. each exam paper had the same probability of receiving a particular ethnicity or gender, and each marker received a near equal proportion of each treatment group) (Abadie, Athey, Imbens, & Wooldridge, 2017). [20]

[20]The use of standard errors is justified in Section A.1.

To summarize our results so far:

Result 1: On average, markers show a small positive bias of 0.06 s.d. towards Chinese names. This result is driven by the exam marks for Business, in which Chinese students get 0.11 s.d. more marks than White students.

Result 2: On average, there is no evidence at all of any difference in the marks awarded to male and female names.

Result 3: On average, there is no evidence of any difference in the marks awarded to Chinese and Adopter names.

3.3 Threshold effects

Although, on average, minimal evidence of racial or gender discrimination has been found so far, perhaps discrimination only exists 'where it matters', i.e. at important mark thresholds. An important threshold at the university where we conducted our study is the pass/fail threshold. Students pass this threshold if they get a score above 50. However, if a student receives a total mark below 50 in a subject, s/he fails and must repeat the subject. Markers may be reluctant to give an exam script a mark below 50, either because they do not want the students to fail or do not want to do any additional work. Examining the distribution of marks at the pass threshold may give us a better indication of the presence of racial or gender bias. For example, if a marker prefers a certain gender or race, they may be more likely to ensure that an exam script with the corresponding gender and/or race passes the threshold compared to an identical exam script with the name from a group they disfavour. Therefore, one race or gender may be more likely to receive a mark just above 50 than another race or gender.

[Figure 3 here]

Figure 3 shows histograms of the mark out of 100 that an exam script received, categorised by the race and gender of the names that were assigned to it. Red lines indicate the cut-off marks for a different grade. Clearly, the 50-mark threshold is important, and we see a large jump in density for all histograms above this threshold (except in the case of Chinese females). Further, there seems to be evidence that this jump varies by race and gender. There is a higher concentration of marks just above 50 for White names, than for Chinese or Adopter names. Similarly, there is a higher concentration of marks just above 50 for scripts with female names than for those with male names. This gender disparity is especially evident

when looking at the marks awarded to White and Adopter female names.[21] These threshold effects are tested more formally in Table 6.

[Table 6 here]

In Table 6, we present the regression results of threshold effects. We exclude the marks for the Business exams from this analysis as our markers did not mark the full exam, meaning they would not know if a student was near a grade threshold.[22] We consider the threshold marks for all the grades available at this university. They are as follows: 50 (Pass), 65 (Credit), 75 (Distinction) and 85 (High Distinction). To do so, we create a binary variable, *JustAbove*, equal to one if a script has a mark between 50 and 54, 65 and 67, 75 and 77, or 85 and 87, (indicating that the script has received a mark that is just above a grade threshold), and zero otherwise. In column (1), we consider the full sample of exam scripts in an OLS model. All coefficients are statistically-insignificant, with the most significant coefficient being that of the Chinese group where exams with Chinese names are around 7.8 percentage points less likely to receive a mark just above a grade threshold compared to a script with a White name.[23]

In order to remove the potential noise from including scripts that may not be close to any grade threshold, we restrict our sample to scripts that received marks between 46 and 54, 63 and 67, 73 and 77, or 83 and 87, in column (2). The dependant variable *JustAbove* is as before. As shown in Column (2), this results in all the coefficients increasing in magnitude and becoming statistically-significant. Exams with Chinese names are 26 p.p. less likely to receive a mark just above a threshold than scripts with White names, whereas exams with Adopter and Female names are 20 p.p. and 19 p.p. less likely, respectively.

Finally, column (3) looks only at the pass/fail (50-mark) threshold. We restrict the sample to exam scripts that received marks between 46 and 54. The dependent variable is equal to one for marks between 50 and 54, and zero for marks between 46 and 49. This is arguably the most economically meaningful threshold, as a student who fails a subject may have to repeat the subject, increasing their financial costs as well as the time they have to spend at university. The results reported in column (3), show a large reduction in the probability of receiving a mark above the pass-fail threshold for scripts with Chinese and Adopter names. The coefficient for the Chinese group is not statistically significant, but this may be a result of the very small sample size of exam scripts in this range. Finally, females are now slightly more

[21] The exact proportion of the distribution with marks between 45-49 and 50-54 is shown in Section B.1

[22] Including the Business marks has the greatest effect on the coefficient of the Chinese group. In column (1), including the Business marks reduces the estimate of this coefficient to -0.002. In column (2), the coefficient becomes -.08 and in column (3) all coefficients are similar.

[23] In a Probit specification of the same model, the coefficient of the Chinese group is larger with an average marginal effect of 14 percentage points and a p-value of 0.105

likely to be advantaged at the 50-mark threshold, but the coefficient is not statistically significant, and is smaller in magnitude compared to the other coefficients.

Result 4: Scripts with Chinese names are less likely to get a mark just above a grade threshold. There is some evidence that scripts with Adopter and female names are also less likely to get a mark above a grade threshold.

Finally, Table 6 shows that there are no significant differences in coefficients between Chinese and Adopter names in all specifications, suggesting that bias affects each group equally, and markers do not distinguish between the two groups.

Result 5: Adopters and Chinese names face the same bias.

This analysis shows that although there is little evidence of any gender or racial bias in marking on average, there is some evidence that bias exists against Chinese, Adopter and female names when grade thresholds are considered in the analysis. Importantly, bias at the thresholds can be of more economic importance than bias at the average. For example, if someone fails to reach the 50-mark threshold, they must repeat the subject. Therefore, bias exists 'where it matters'.

Finally, the experiment also included four markers who marked all their exam scripts with no names on the exam cover sheets. The results including these four markers are broadly consistent with the results already reported. This analysis can be found in Appendix Section A.5 for average effects and Section B.2 for threshold effects.

4 Interpreting the Results

We now turn to interpreting the results, focusing, in particular, on whether they are consistent with taste-based or statistical discrimination. To summarize, the results show a slight positive bias towards exams with Chinese names, which is driven by the marks for the Business subject. Further, it was found that Chinese students may be disadvantaged at the grade thresholds. In this section, we further discuss these results, and provide additional validation through administrative data and power calculations. Finally, we discuss the limitations of this study.

4.1 Statistical vs. Taste-based discrimination

As noted in Section 1, the two main theories in the Economics literature of discrimination are that of taste-based and statistical discrimination. Ultimately, the experimental results contain elements of discrimination that could align with both types of discrimination. The result that the probability of getting a mark just above a threshold differs by race is more consistent with taste-based discrimination. The fact that this is only seen in the vicinity of grade thresholds but not on average suggests that giving an extra mark is an explicit decision made by the marker. In contrast, if this result had been consistent with statistical discrimination, any bias would have occurred throughout the mark distribution, not just close to a grade threshold.

However, our results also show that Chinese students have an advantage at the average (primarily driven by the marks for the Business subject). This is consistent with statistical discrimination given Chinese students have the highest average marks in the business subject compared to the other two subjects.[24] Further, in Section A.3 of the Appendix, we examine whether bias varies by the amount of grading experience a marker has. If marking is more consistent with taste-based discrimination, a marker would demonstrate an aversion to a different ethnicity or gender, that should be independent of their prior marking experience. However, if marking is consistent with statistical discrimination, then markers would use their beliefs about group characteristics to supplement the information on an exam script when grading it. Here, bias may differ by marking experience, as a marker's beliefs of group characteristics may change with their marking experience. Table A1 shows evidence that discrimination may vary by marker experience which is consistent with statistical discrimination. Further, we find limited evidence of in-group bias (indicative of taste-based discrimination) when examining the results in Table A2 in Appendix Section A.4. However, both these results may be influenced by the small number of markers in our experiment, as discussed further in Section 4.3.2.

In sum, the strongest evidence shows that threshold bias is more consistent with taste-based discrimination, whereas the bias at the average is more consistent with statistical discrimination.

4.2 Evidence from Administrative Data

We received de-identified students' marks for the three subjects considered in our experiment for the years 2014 to 2018 from the university. The student names associated with these marks were coded as Chinese

[24]The higher average marks in Business subjects obtained by Chinese students compared to White students is presented in section 4.2, particularly in 7.

or White, allowing us to examine the average marks associated with the three main name groups in our experiment: White, Chinese and Adopter. Data was also provided about the gender of the students. The university we conducted our experiment in adopted anonymous marking in 2018, meaning that markers no longer saw names on exams from then on. Therefore, examining how the trend in marks differed in 2018 compared to prior years may give insights into the validity of our experimental results.

[Figure 4 here]

Figure 4 examines how marks have changed in the last five years in the three subjects included in the experiment. If our experimental results are consistent with actual marking behaviour, then we would expect the marks for Chinese students to have worsened compared to those of White students following the introduction of anonymous marking in 2018. Figure 4 shows that there was a statistically significant decline in the marks of students of all racial groups from 2014 to 2018. Therefore, it is hard to ascertain how marks changed in 2018 following the introduction of anonymous marking, as there is no counterfactual as to how marks would have changed in 2018 in the absence of anonymous marking. For White and Chinese students, the change in marks from 2017 to 2018 is above their respective trend. On average, marks for White students declined by 1.21 marks year-on-year between 2014 and 2017, but increased by 0.69 marks between 2017 and 2018. The marks for Chinese students decreased by 2.71 marks year-on -year between 2014 and 2017, but only by 2.09 marks between 2017 and 2018. The greater increase in marks for White students compared to that of Chinese students is broadly consistent with our experimental results, which suggest that having names on an exam benefits those with Chinese names. However, Adopters saw a greater decline between 2017 and 2018, though this was still smaller than the decline between 2015 and 2016.

[Figure 5 here]

Figure 5 focuses solely on administrative data for the Business subject, as the existence of a positive bias towards Chinese names at the average is driven by the marks for the Business paper. Figure 5 shows that there was an increase in marks in 2018 compared to 2017. However, all racial groups had an almost identical increase. If our experimental results are consistent with this data, we would expect there to have been bias in favour of the Chinese students in the years with non-anonymous marking, and the increase in marks seen across the board in 2018 to have been smaller for Chinese students than for White or Adopter students. However, these plots are only indicative of what the effects of anonymous marking could be, and cannot account for the many other factors that have contributed to the general decline in course marks since 2014.

[Figure 6 here]

The administrative data also sheds some light on why the positive bias towards Chinese names is driven by the marks for Business. Figure 6 graphs the proportion of Chinese students enrolled in our three subjects from 2014 to 2018. The proportion of students with Chinese names increased for all subjects, but the increase was much greater for Economics and IT, where it almost doubled, compared to Business which saw a less than 10 percentage point increase. If discrimination is theorised as a process where markers learn from the past performance of particular racial groups, then markers for Business would have a much better idea of the performance of Chinese students than markers for either Economics or IT. This is because the exposure of markers to Chinese students would have stayed relatively constant in Business, given the limited variation in the proportion of Chinese students enrolled in the subject from 2014 to 2018. In contrast, the Economics and IT markers may not have extensive and consistent prior experience with the performance of Chinese students, given the proportion of Chinese students in these two subjects increased dramatically from 2014 to 2018. As such, they may not have yet formed an opinion about the general performance of Chinese students which could bias their marking. Consequently, if bias is consistent with statistical discrimination, then any bias towards Chinese students would likely only occur in Business, where markers have already formed an opinion about their performance, but not in Economics and IT, where markers may still be unfamiliar with the performance of Chinese students. This is consistent with experimental results.[25]

[Table 7 here]

Further, if students with Chinese names do relatively well in Business, this could confirm our argument for the existence of statistical discrimination. Table 7 shows the average marks different racial groups received in the three subjects. All differences in Table 7 are statistically significant. As can be seen, Business is the only subject where Chinese students on average received a mark higher than White students. Therefore, our experimental results are consistent with the presence of statistical discrimination. Namely, the overall positive bias towards Chinese students driven by the marks in the Business exam arises because of the markers' pre-conceived beliefs about their performance.

[25] Between 2014 and 2018, the growth in enrolment for each subject was different. Even so, the enrolment patterns confirm our argument. Enrolment in IT and Economics increased by 97% and 37%, respectively, between 2014 and 2018, meaning there would have been a large influx in both the number and proportion of Chinese students in these subjects over this period. In contrast, the change in the number of students in Business was significantly smaller, growing by a modest 22%.

4.3 Potential Limitations

4.3.1 Minimum Effect Sizes

One possible limitation of our experiment is the small sample size, which affects the extent to which we can draw conclusions about our main research questions.

Firstly, the average estimated mark differences among White, Chinese and Adopter names in our preferred model in column (4) of Table 4 are below our detectable effect size. This is true of all the coefficients in Table 5, except for the coefficient of the Chinese group in the Business subject, which is statistically significant. Our experiment has the power to detect an effect size of around 1.5 marks or 0.07 s.d in the full sample and of around 2 marks or 0.1 s.d for each subject. This means that if quantitatively small amounts of racial bias are present in marking, this particular experiment does not have sufficient power to be able to detect it.

However, our experiment does have the power to detect effect sizes that have been found in similar prior experimental studies. For example, Hanna and Linden (2012) found effect sizes from 0.03 to 0.08 s.d. in their study and Sprietsma (2013) found effect sizes of 0.1 s.d. Further, it is unlikely that bias below 0.07 s.d, which is equivalent to 1.5 marks out of 100, has any practical implications for student welfare or university management. For example, the average standard deviation of marks given to the same exam script across all the subjects is 8.8 marks. Therefore, racial bias below 1.5 marks would be minimal in comparison to the normal marking error given for the exact same exam paper. In fact, our power calculations give more weight to the experimental result that Chinese names get a mark bonus of 0.06 s.d., which, at a p-value of 0.105, is quite significant for such a small effect size.

Moreover, the same argument applies when considering the estimated difference in marks between male and female names. However, the effect sizes found in the literature are much larger. For example, Lavy (2008) found differences of up to 0.272 s.d. between males and females, whereas Breda and Ly (2015) found differences of 0.25 s.d. Effect sizes of these magnitudes would have been detected in our experiment had they existed, increasing our confidence in the result that there is no difference in marks awarded to male and female names on average.

4.3.2 Small Number of Markers

Given our experiment had 14[26] unique markers marking exams for three subjects, the external validity of the marks rewarded is a potential concern. It is unclear whether the biases of these markers represent the biases of all markers in general, or if they are just specific to these 14 markers. For example, 5 markers were responsible for grading the Business papers. Given the overall positive bias towards Chinese names is driven by the Business marks, a key question would be if the positive bias is simply due to the 5 different markers, as opposed to the different context of grading. However, as discussed in the previous section,this positive bias can also be explained by the difference in the proportion of Chinese students enrolled in Business compared to IT or Economics, and their average performance in the subject. Furthermore, our threshold results do not seem to be driven by any individual subject.

5 Conclusions

University marks have an important impact on outcomes in the labour market, meaning that examining potential discrimination in university marking can provide insights into consequent labour market inequalities. This paper examined the existence of race or gender discrimination in university marking. On average, we found that exams with Chinese names are given a practically small marking bonus of 0.06 s.d. This result is consistent with statistical discrimination. Notably, no other statistically-significant differences between genders, or between Adopter, Chinese and White names are observed.

However, we also find that bias exists at grade thresholds, with exams with Chinese and Adopter names less likely to receive a mark above a threshold. These results are consistent with taste-based discrimination and suggest that discrimination exists 'where it matters'. The results also demonstrate that markers don't distinguish between Chinese and Adopter names.

Our findings provide ample directions for future research. Firstly, it is likely that the scope for potential discrimination at university is broader than discriminatory marking alone. Certain groups might face bias prior to being assessed, which in turn could affect the quality of their work. For example, a biased teacher may offer less assistance and support to students from a group they dislike, which could affect a student's understanding and/or confidence. Further research should thus be undertaken about how other factors at university could affect marking discrimination.

Secondly, even though our paper finds some evidence of bias against Chinese and Adopter students, it

[26]The results from the four additional markers who marked unnamed exams are reported in the Appendix.

is not large enough to explain the significant gap in marks that we find in administrative data between students with White and Chinese names. This suggests that studying the potential source of this inequality can be a fruitful area for future research, and might include investigating the cause/s of the overall negative trend in marks awarded since 2014.

These results also have significant policy implications. Anonymous marking is one potential policy which can reduce the biases we found in our experiment in a relatively costless way. This small modification to the way marking is done can mean a fairer process, and is a policy that can be easily implemented at universities around the world.

Further, our finding that marking bias against Adopters exists at grade thresholds is significant, as Adopters tend to be domestic students who go on to seek employment in the Australian labour market after they graduate. Thus, the labour market inequalities between White-Australians and Chinese-Australians might be partially explained by discriminatory marking at the university level.

However, anonymous marking should not be seen as a panacea for reducing racial and/or gender inequalities in the labour market, or even for countering the disparity in marks given to White and Chinese students at universities. The biases we found in our experiment were relatively small, and on average, actually mean that Chinese students are expected to fare worse under anonymous marking given the small positive bias towards them. As such, a broad range of policies would need to be adopted to tackle such labour market inequalities.

References

Abadie, A., Athey, S., Imbens, G., & Wooldridge, J. (2017, 10). *When Should You Adjust Standard Errors for Clustering?* (Tech. Rep.). Retrieved from `http://arxiv.org/abs/1710.02926` `http://www.nber.org/papers/w24003.pdf` doi: 10.3386/w24003

American Psychological Association. (2016). *Stress in America: The impact of discrimination* (Tech. Rep.). Retrieved from `www.stressinamerica.org`

Arrow, K. J. (1972). Some mathematical models of race discrimination in the labor market. *Racial discrimination in economic life*, 187–204.

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

Bertrand, M., Chugh, D., & Mullainathan, S. (2005). Implicit discrimination. *American Economic Review*, *95*(2), 94–98. doi: 10.1257/000282805774670365

Bertrand, M., & Duflo, E. (2017). Field Experiments on Discrimination. In *Handbook of field experiments* (Vol. 1, pp. 309–393). doi: 10.1016/bs.hefe.2016.08.004

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal ? A Field Experiment on Labor Market Discrimination. *American Economic Associaton*, *94*(4), 991–1013.

Breda, T., & Ly, S. T. (2015). Professors in core science fields are not always biased against women: Evidence from France. *American Economic Journal: Applied Economics*, *7*(4), 53–75. doi: 10.1257/app.20140022

Chowdhury, S., Ooi, E., & Slonim, R. (2020). Racial Discrimination and White First Name Adoption : Evidence from a Correspondence Study in the Australian Labour Market. *IZA Discussion Paper*, *13208*.

Feld, J., Salamanca, N., & Hamermesh, D. S. (2016). Endophilia or Exophobia: Beyond Discrimination. *Economic Journal*, *126*(594), 1503–1527. doi: 10.1111/ecoj.12289

Floyd, et al. v. City of New York, et al (Vol. 1:08-cv-01). (2013).

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

Hanna, R. N., & Linden, L. L. (2012). Discrimination in grading. *American Economic Journal: Economic Policy*, *4*(4), 146–168. doi: 10.1257/pol.4.4.146

Hartocollis, A. (2019). *Harvard Does Not Discriminate Against Asian-Americans in Admissions, Judge Rules.* Retrieved from `https://www.nytimes.com/2019/10/01/us/harvard-admissions-lawsuit.html`

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evi-

dence from a natural experiment. *Journal of Public Economics*, *92*(10-11), 2083–2105. doi: 10.1016/j.jpubeco.2008.02.009

Mujcic, R., & Frijters, P. (2013). Still Not Allowed on the Bus: It Matters If You're Black or White! *IZA Discussion Paper*(7300).

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature*, *56*(3), 799–866. doi: 10.1257/jel.20161309

Niederle, M., & Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: The Role of Competition. *Journal of Economic Perspectives*, *24*(2), 129–144. Retrieved from http://pubs.aeaweb.org/doi/10.1257/jep.24.2.129 doi: 10.1257/jep.24.2.129

Phelps, E. S. (1972). American Economic Association The Statistical Theory of Racism and Sexism. *Source: The American Economic Review*. doi: 10.2307/1806107

Presidential Task Force on Inclusion and Belonging. (2018). *Pursuing Excellence on a Foundation of Inclusion* (Tech. Rep.). Harvard University. Retrieved from https://inclusionandbelongingtaskforce.harvard.edu/files/inclusion/files/harvard$_i nclusion_b elonging_t a$

Race Equality Charter. (2020). Retrieved from https://www.ucl.ac.uk/equality-diversity-inclusion/equality-area

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school teachers. *Empirical Economics*, *45*(1), 523–538. doi: 10.1007/s00181-012-0609-x

Van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, *30*(5), 1045–1058. doi: 10.1016/j.econedurev.2011.05.008

6 Tables & Figures

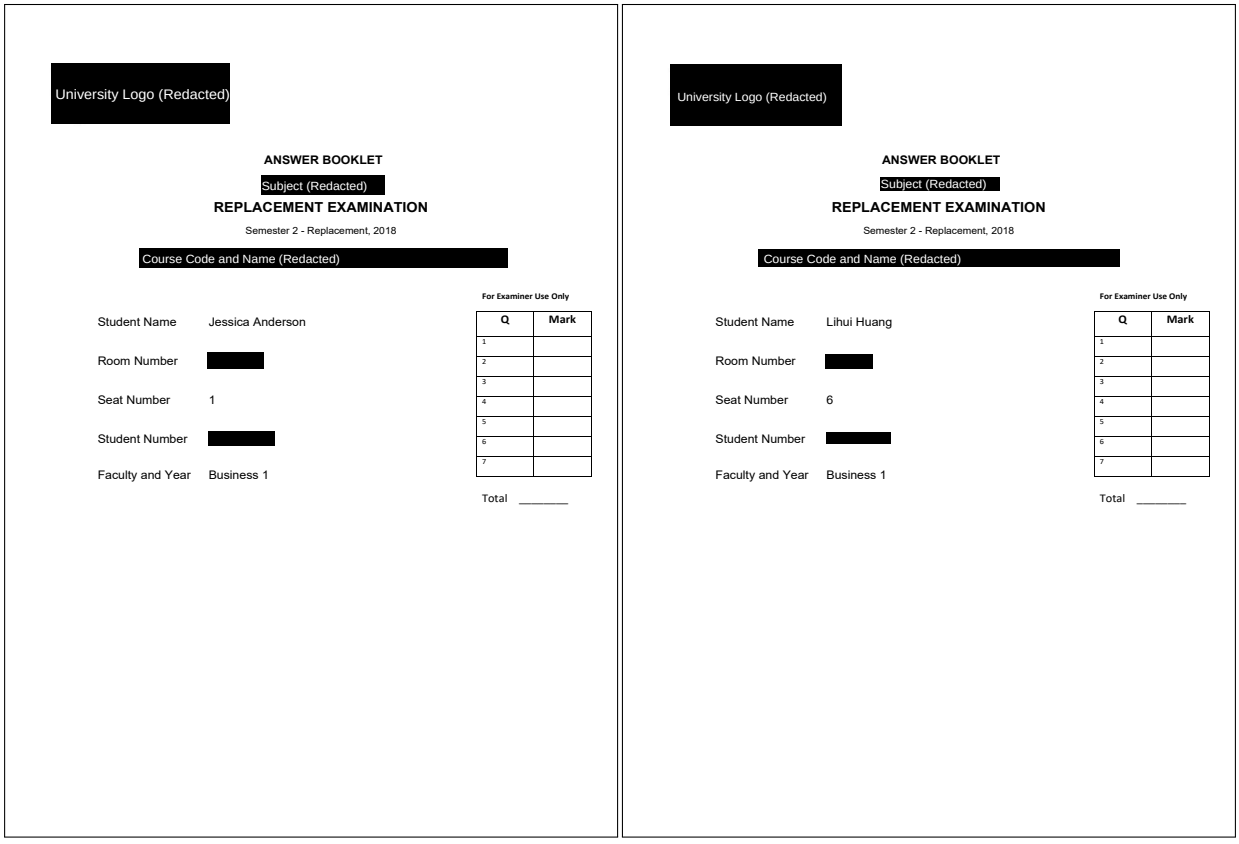

Figure 1: Left: A typical coversheet placed on an exam script of a Business, with a White name. Right: A typical coversheet placed on a Business exam with a Chinese name. Coversheets closely mimic the regular coversheets used on the respective exams. Black boxes are placed on sensitive information that is unable to be disclosed.

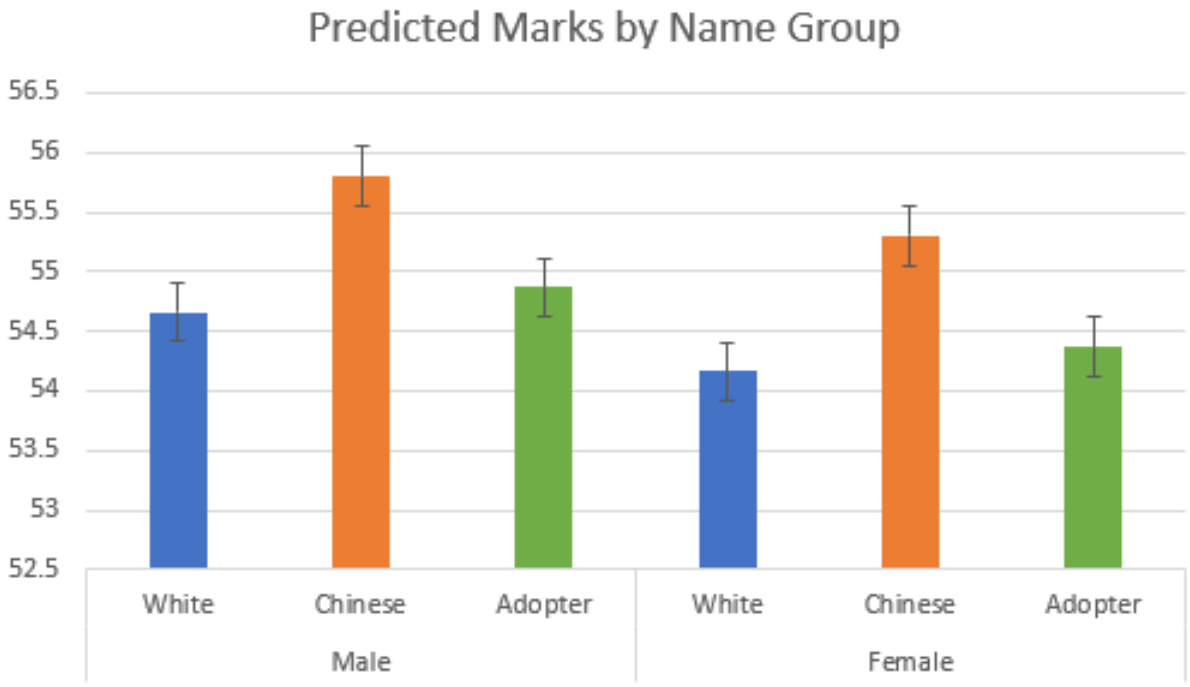

Figure 2: Predicted Average Marks

Notes: Predicted Average Marks from column (4) of Table 4. Error bars refer to one standard error of the ethnicity coefficients.

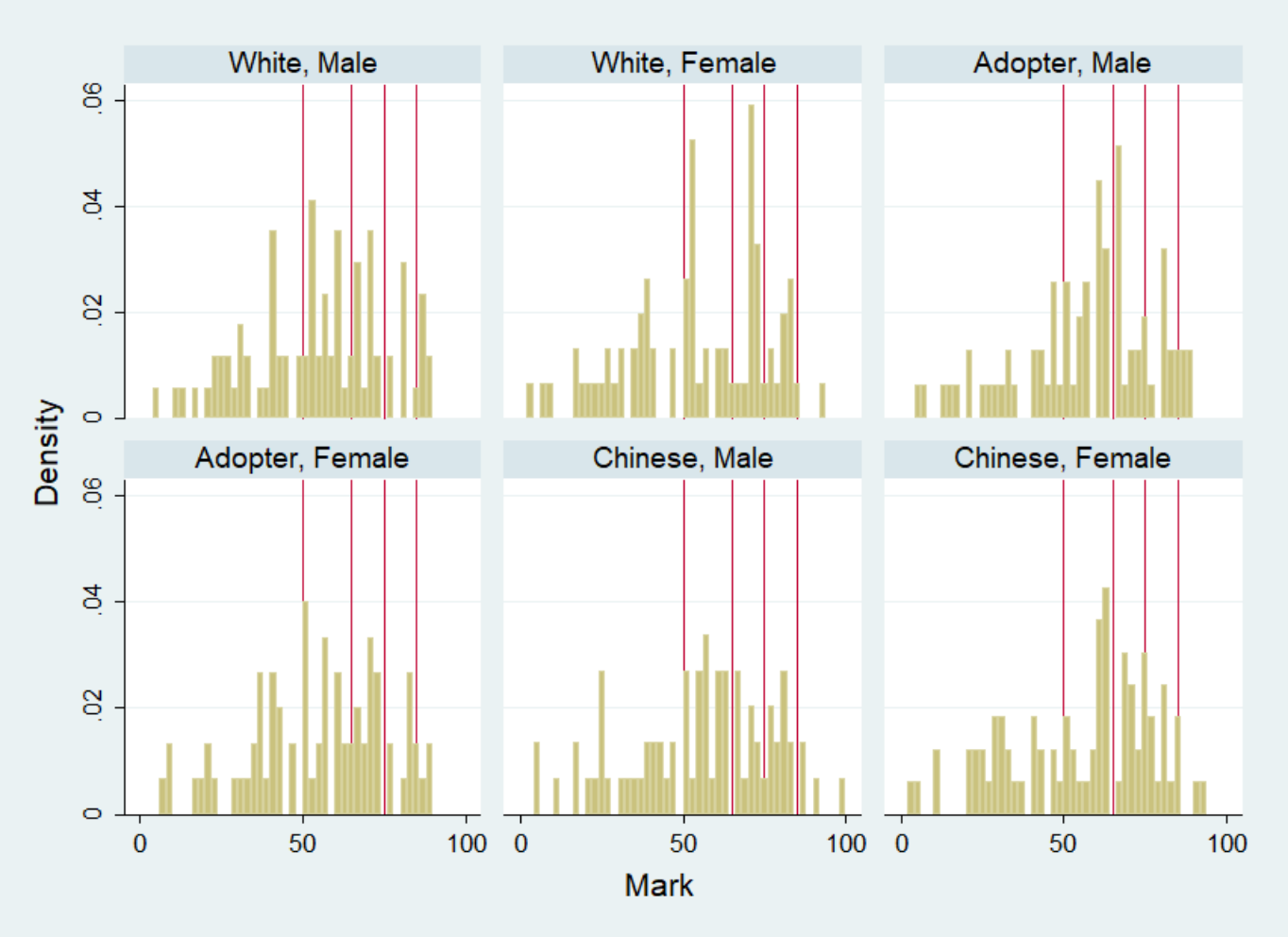

Figure 3: Histograms of Mark by Race and Gender

Notes: Histogram of standardised marks out of 100 for pooled sample of all subjects. Red lines indicate marks where the grade changes.

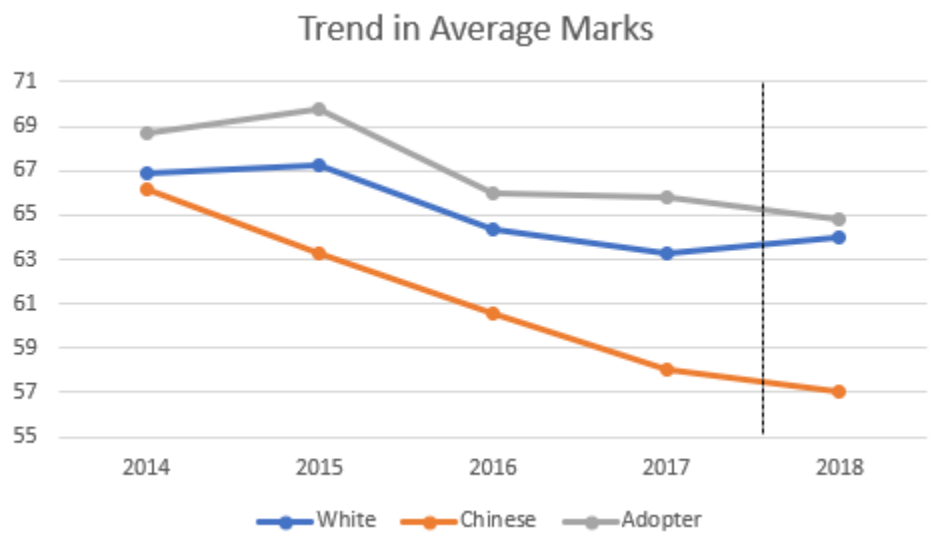

Figure 4: Average mark by year for each racial group

Note: Name Groups are unweighted. Anonymous marking began in 2018, shown by dotted line.

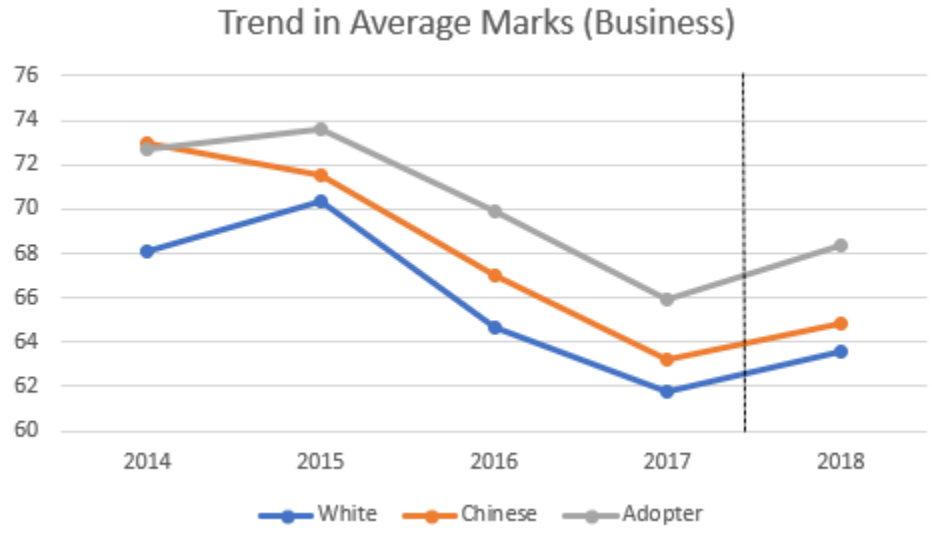

Figure 5: Average mark by year for each racial group (Only Business)

Note: Name Groups are unweighted. Anonymous marking began in 2018, shown by dotted line.

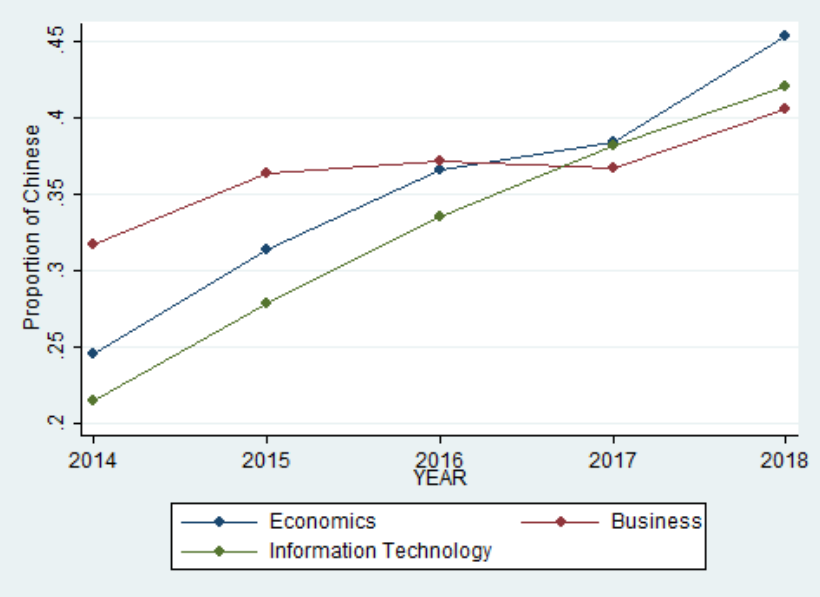

Figure 6: Proportion of students with Chinese names by subject

Table 1: Number of markers and exam scripts per condition

	Economics	Business	IT
Markers	5	5	4
Scripts	30	36	35

Table 2: Proportion of Ethnicities

	Experiment			Official Names			Preferred Names		
	Economics	Business	IT	Economics	Business	IT	Economics	Business	IT
White Male	0.19	0.18	0.17	0.18	0.17	0.19	0.18	0.17	0.20
White Female	0.15	0.16	0.17	0.092	0.093	0.055	0.094	0.097	0.056
Chinese Male	0.20	0.13	0.14	0.18	0.14	0.27	0.10	0.052	0.16
Chinese Female	0.14	0.19	0.19	0.28	0.27	0.15	0.14	0.088	0.068
Adopter Male	0.17	0.17	0.16	0.035	0.050	0.070	0.12	0.14	0.19
Adopter Female	0.15	0.17	0.16	0.024	0.054	0.025	0.16	0.24	0.11
Other Male	0	0	0	0.12	0.14	0.17	0.12	0.14	0.15
Other Female	0	0	0	0.089	0.086	0.069	0.085	0.075	0.064

Table 3: Summary Statistics of Marker Characteristics

	Economics	Business	IT
Experience	2.40	5.60	5
	(2.702)	(7.829)	(6)
White	0.20	0.60	0.50
	(0.447)	(0.548)	(0.577)
Chinese or Adopter	0.60	0	0.50
	(0.548)	(0)	(0.577)
Other	0.20	0.40	0
	(0.447)	(0.548)	(0)
Female	0.20	0.20	0.25
	(0.447)	(0.447)	(0.500)
Observations	5	5	4

Note: Shown is the mean of a variable for a subject with standard deviations in parentheses. Experience refers the number of semesters graders have marked any subject prior to the experiment. All other variables refer to the ethnicity of the grader, which is ascertained from the name of the grader. "Other" is equal to one if graders have a different ethnicity to the ethnicities on the list.

Table 4: Difference in Marks Based on Race and Gender - Average Effects

	(1) Mark	(2) Mark	(3) Mark	(4) Mark
Chinese	0.948	0.975	0.927	1.140
	(2.384)	(2.303)	(0.892)	(0.702)
Adopter	1.674	2.282	0.153	0.206
	(2.394)	(2.316)	(0.904)	(0.708)
Female	-0.758	-0.0899	-0.326	-0.501
	(1.958)	(1.882)	(0.713)	(0.568)
chinesemarker		0.723	0.624	
		(3.736)	(1.385)	
femalemarker		-1.737	-1.414	
		(2.605)	(0.966)	
adoptermarker		2.129	2.387*	
		(3.778)	(1.401)	
othermarker		2.140	2.271**	
		(2.756)	(1.022)	
Experience		-0.421**	-0.418***	
		(0.190)	(0.0706)	
Constant	54.38***	55.69***	68.53***	73.17***
	(1.909)	(4.482)	(3.597)	(2.890)
Subject Fixed Effects	No	Yes	No	No
Script Fixed Effects	No	No	Yes	Yes
Grader Fixed Effects	No	No	No	Yes
R squared	0.0014	0.10	0.90	0.94
Observations	470	469	469	470
Number of Exam Scripts	101	101	101	101
Test of Adopter=Chinese	0.76	0.57	0.39	0.19

Note: OLS regressions; The dependent variable is the standardised mark an exam script received out of 100. Each model is an OLS regression on a pooled sample of all subjects. In Coloumns (2) and (3) additional controls include order returned and its square. (Standard errors are not clustered.) * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Direct Bias (By Subject)

	Economics	Business	IT
	(1)	(2)	(3)
	Mark	Mark	Mark
Chinese	1.126	2.246**	-0.407
	(1.505)	(1.009)	(1.111)
Adopter	1.113	0.765	-1.610
	(1.546)	(1.005)	(1.110)
Female	-0.774	-0.761	-0.0195
	(1.250)	(0.835)	(0.850)
Constant	72.91***	74.01***	52.71***
	(3.659)	(2.682)	(2.684)
Script Fixed Effects	Yes	Yes	Yes
Grader Fixed Effects	Yes	Yes	Yes
R squared	0.82	0.95	0.97
Observations	150	180	140
Test of Adopter=Chinese	0.99	0.15	0.29

Note: OLS regressions; The dependent variable is the standardised mark an exam received out of 100. Each regression is on one subject, indicated in the first row of the table. (Standard errors are not clustered.)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Threshold Effects

	(1) P(Just Above)	(2) P(Just Above)	(3) P(Above 50)
Chinese	-0.0780	-0.261**	-0.285
	(0.0602)	(0.107)	(0.283)
Adopter	0.0109	-0.195*	-0.792**
	(0.0611)	(0.109)	(0.240)
Female	-0.00910	-0.193**	0.0861
	(0.0480)	(0.0888)	(0.363)
Constant	0.201	0.433*	1.660**
	(0.194)	(0.227)	(0.588)
Script Fixed Effects	Yes	Yes	Yes
Grader Fixed Effects	Yes	Yes	Yes
R squared	0.35	0.72	0.90
Observations	290	109	41
Number of Exam Scripts	65	46	26
Test of Adopter=Chinese	0.15	0.54	0.19

Note: All models are OLS specifications. For models (1) and (2), the dependent variable is equal to one if the mark is between 50-54, 65-67, 75-77 or 85-87, and zero otherwise. For model (2), the dependent variable is equal to one if a mark is between 50 and 54, and zero otherwise. The full sample of marks is shown in columns (1), marks between 46-54, 63-67, 73-77 and 83-87 in columns (2), and marks between 46 and 54 in column (3). (Standard errors are not clustered).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Average Mark by Racial Group

	Economics	Business	IT
White	64.1	65.4	66.3
Adopter	66.5	69.7	62.6
Chinese	57.1	67.3	52.7
Other	61.0	64.2	58.2

Note: The Average mark is a standardised mark out of 100. All mark differences between name groups are statistically-significant.

A Appendix

A.1 Standard Errors

Following Abadie et al. (2017), in regressions with script and marker fixed effects, we do not use clustered standard errors, as clustering standard errors when there are fixed effects, are only necessary if there is either clustering in the sampling process and heterogeneity in the treatment effect, or clustering in the assignment mechanism as well as heterogeneity in the treatment effect. Assignment to treatment is not clustered, as each treatment (i.e. name group) is given to each marker. However, when looking at the effects of anonymous exams (Section A.5), the treatment of marking anonymous exams is clustered around the marker, as either a marker marks exams that are all anonymous or all have names on them. Therefore, these regressions will have their standard errors clustered at the marker level.

A.2 Power Simulations

As our experiment has fixed effects on two dimensions, the exam script and marker dimensions, we attempt to simulate statistical power.

We first make some simplifications. We assume that there are only two treatment groups, W and C, and calculate the power to detect an x unit increase due to having a name C on the exam paper, varying x (the effect size), s the number of exam scrips and m the number of markers. The simulation proceeds in the following steps:

1. Randomly select s number of marks,[27] and group them into three groups of $s/3$, representing the three subjects in the experiment
2. These three groups marks represent three markers grading in three different subjects. To create more markers we duplicate these original marks ($Mark$) by adding random noise specific to a marker m, to create a new mark $NewMark$, such that:
    - $NewMark_{sm} = Mark_{sm} + \epsilon_m$
    - ϵ_m refers to a marker specific random noise $\epsilon_m \sim N(0, 3.33)$[28]

[27] Selected from the dataset of the experimental results.

[28] The standard error is calculated from the results. It is the standard error associated with the average standard deviation of exam scripts for the same marker.

3. Like in the experiment, each exam paper randomly gets assigned a "name" W or C. Each marker "marks" all the scripts for one "subject". Each unique script alternates between W or C depending on the marker

4. Each mark is replaced again. If the group is W:

$$FinalMark_{sm} = NewMark_{sm} + \gamma_{sm}$$

   If the group is C:

$$FinalMark_{sm} = NewMark_{sm} + x + \gamma_{sm}$$

   x represents the effect size or the bias from having C on the exam paper, and γ represents the random noise generated from an exam script, where $\gamma_{sm} \sim N(0, 3.45)$[29]

5. Run the regression:

$$Mark_{sm} = \beta_0 + \beta_1 C + \alpha_s + \alpha_m + \epsilon_{sm}$$

   where α_s and $alpha_m$ are script and marker fixed effects respectively

6. Repeat steps (1) - (5) 1000 times, saving the p-value of β_1. Calculate power as the proportion of p-values that are less than or equal to 0.05 out of 1000

The results are shown in figures 7 and 8 . Figure 7 shows power to detect effect sizes of $x = 1, 1.5, 2$ marks out of 100, given a varying number of markers. The number of exam scripts are held constant at 30 for each subject, meaning 90 in total. For example, in figure 7, an experiment with 2 markers in each subject marking 30 exams each, meaning 6 markers in total, will have the power to detect a 2 mark difference with power level of 0.83. Figure 7 shows that an experiment with 12 markers would be able to detect a one mark effect with more than 0.8 power, while an experiment with 6 markers would be able to detect a 1.5 mark effect with more than 0.8 power.

Figure 8 looks at how power changes with varying numbers of unique exam scripts. The number of markers is held constant at four per 30 exam scripts. Figure 8 shows that with 30 exam scripts a 2 mark effect size can be detected with more than 0.8 power.

These power results are relevant for our experiment. Although there can be up to five markers grading the exact same exam script in our experiment, there are six possible treatment groups. Therefore, when comparing two groups like White and Chinese, for the exact same exam script there will be at most two

[29]The standard error is calculated from the results. It is the standard error associated with the average standard deviation of each unique exam script across markers.

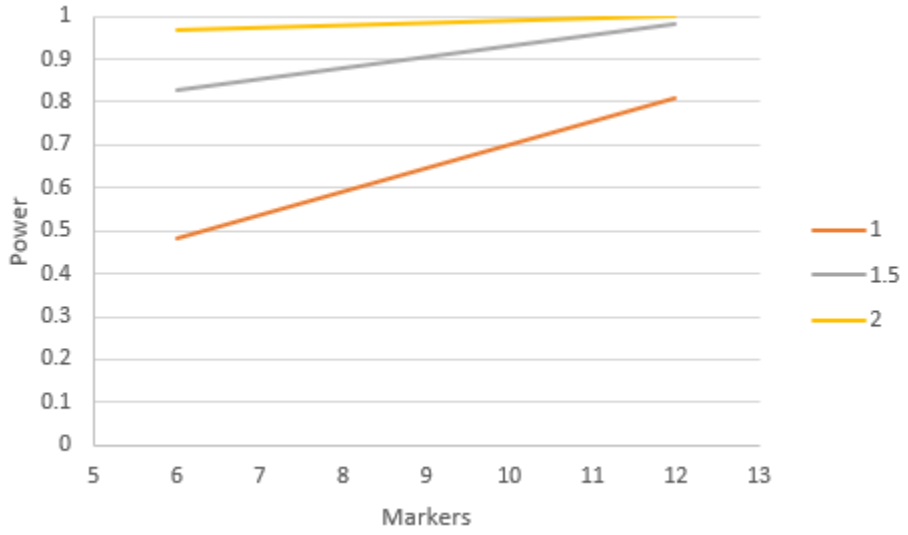

Figure 7: Power at different effect sizes and markers

Note: Different lines refer to the respective effect sizes out of 100. The number of exam scripts is held constant at 90, which is 30 per subject. Power is only calculated at Markers= 6 and Markers= 12

markers grading the script with a White name, and at most two more markers grading the same script with a Chinese name. Even so, our power calculations have shown that we can comfortably detect an effect size of 1.5 marks when looking at all subjects, or 2 marks when looking at any individual subject, which typically has 30 exam scripts.

A.3 Mechanisms

As discussed in Section 1, discrimination is theorised in terms of statistical discrimination and taste-based discrimination. One way to disentangle the effects of the two, is to examine if discrimination varies by the marking experience of the marker. This is defined as how many semesters a marker has previously provided marking assistance for any subject. If marking were more consistent with taste-based discrimination, a marker should have an aversion to a different ethnicity or gender, that should be independent of prior marking experience. However, if marking is consistent with statistical discrimination, then markers will use their beliefs of group characteristics to supplement the information on an exam script when giving a mark. Here, bias may differ by marking experience, as markers beliefs of group characteristics would change with marking experience. For example, a less experienced marker is more likely to rely on group stereotypes, whereas a more experienced marker may be more likely to rely on their prior experience of how different groups have performed in the past.

Table A1: Interactions with Marker Experience

	All Subjects	Economics	Business	IT	All Subjects	Economics	Business	IT
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mark	Mark	Mark	Mark	Mark	Mark	Mark	Mark
Chinese	0.567	-0.718	2.656**	-3.252**	0.452	-0.913	2.464**	-3.111***
	(0.903)	(2.339)	(1.296)	(1.494)	(1.078)	(1.633)	(0.634)	(0.179)
Adopter	1.074	0.0600	2.592**	-1.836	1.217	0.174	2.757**	-1.987**
	(0.906)	(2.357)	(1.272)	(1.495)	(1.002)	(2.678)	(0.809)	(0.444)
Female	0.0831	-0.817	0.551	0.696	0.508	-0.843	1.229	0.567
	(0.847)	(1.863)	(1.438)	(1.272)	(1.244)	(2.838)	(1.703)	(0.375)
Adopter × Experience	-0.231*	0.422	-0.348**	-0.0562	-0.247**	0.340	-0.378**	-0.0410
	(0.135)	(0.715)	(0.150)	(0.222)	(0.102)	(0.662)	(0.129)	(0.0936)
Chinese × Experience	0.105	0.751	-0.104	0.538**	0.114	0.769	-0.102	0.533***
	(0.133)	(0.722)	(0.150)	(0.217)	(0.110)	(0.490)	(0.0710)	(0.0283)
Female × Experience	-0.116	0.0577	-0.185	-0.119	-0.165	0.0732	-0.224	-0.107
	(0.132)	(0.585)	(0.178)	(0.192)	(0.132)	(0.506)	(0.150)	(0.0501)
Experience					-0.309*	0.796	-0.317	-0.566*
					(0.165)	(0.556)	(0.229)	(0.195)
Constant	73.33***	71.25***	72.86***	52.93***	68.91***	66.66***	68.13***	56.19***
	(2.883)	(4.285)	(2.858)	(2.755)	(5.821)	(6.402)	(3.509)	(4.427)
Script Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grader Fixed Effects	Yes	Yes	Yes	Yes	No	No	No	No
R squared	0.94	0.82	0.95	0.97	0.90	0.75	0.92	0.97
Observations	470	150	180	140	470	150	180	140
Test of Adopter=Chinese	0.58	0.75	0.96	0.35	0.29	0.50	0.79	0.14

Note: The dependent variable is the standardised mark an exam received out of 100. Columns (1) - (4) have both Script and Grader fixed effects, while Columns (5) - (8) only have script fixed effect. Standard Errors for Columns (1) - (4) are not clustered, while for Columns (5) - (8) are clustered by grader

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

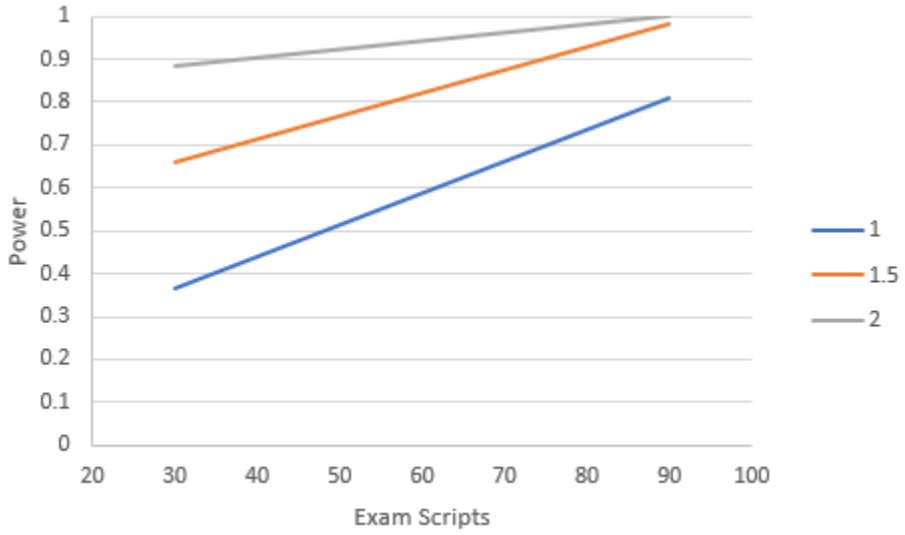

Figure 8: Power at different effect sizes and scripts

Note: Different lines refer to the respective effect sizes out of 100. The number of markers is held constant at 4 per 30 exams.Power is only calculated at Scripts= 30 and Scripts= 90

Table A1 examines this hypothesis, and finds some evidence that bias varies by marker experience. Columns (1) - (4) are regressions with exam script and marker fixed effects and add interaction terms where Chinese, Adopter and Female are interacted with marker experience. The regressions are for all subjects, Economics, Business, and IT, respectively. Columns (5) - (8) show the same regressions, but only have exam script fixed effects to include the variable experience in the regression, and standard errors are clustered at the marker level. In column (1), an additional semester of marker experience reduces the mark for an exam with an adopter name by 0.23 marks compared to a white male, while in the Business subject the decrease is 0.35 marks. In the IT subject, in column (3), a marker with an additional semester of experience increases the mark they give an exam with a Chinese name by 0.54 marks. Results are similar in models (5) - (8), where in column (5) every additional semester of marking experience for a marker, reduced the mark they give by 0.31 marks. An analysis of how discrimination differs by marker ethnicity and a possible relation to taste-based discrimination is examined in the next section

A.4 Marker Ethnicity

Section 4.1 examined how discrimination varied by marker experience, but another potential mechanism through which discrimination could occur is in-group bias (Hanna and Linden (2012)). In Table A2 two variables are added to the models of Table 5. 'Race Match' is equal to one if the ethnicity (White, Chinese

Table A2: Interactions with Marker Ethnicities

	All Subjects	Economics	Business	IT	Economics	Business	IT
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mark	Mark	Mark	Mark	Mark	Mark	Mark
Chinese	1.184	-0.319	1.163	-0.498	2.959*	2.439	-2.061
	(0.724)	(1.680)	(1.644)	(1.247)	(1.634)	(1.754)	(1.437)
Adopter	0.265	2.016	-0.429	-1.561	1.060	0.776	-1.615
	(0.745)	(1.591)	(1.748)	(1.128)	(1.507)	(1.020)	(1.093)
Female	-0.421	-1.666	-0.279	0.376			
	(0.731)	(1.595)	(1.111)	(1.050)			
Race Match	0.209	3.840*	-1.901	-0.187			
	(0.791)	(2.033)	(2.274)	(1.233)			
Gender Match	0.126	-1.559	0.755	0.779			
	(0.784)	(1.789)	(1.114)	(1.183)			
Chinese × WhiteMarker					-8.993**	-0.449	3.731*
					(3.596)	(2.284)	(2.107)
Constant	73.11***	74.87***	73.81***	52.34***	73.45***	73.45***	53.29***
	(2.942)	(3.811)	(2.981)	(2.843)	(3.547)	(2.715)	(2.627)
Script Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grader Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R squared	0.94	0.83	0.95	0.97	0.83	0.95	0.97
Observations	470	150	180	140	150	180	140
Number of Exam Scripts							
Test of Adopter=Chinese	0.20	0.23	0.12	0.43	0.26	0.31	0.76

Note: The dependent variable is the standardised mark an exam received out of 100. Race [Gender] Match is when a grader's race [gender] matches with the marker's. Standard errors are not clustered.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

or Adopter) on the exam script matches with the marker's ethnicity, and 'Gender Match' is equal to one if the gender of the exam script matches the gender of the marker. Column (1) examines this model for all subjects, whereas columns (2), (3), and (4) restrict the samples to Economics, Business and IT respectively. There is limited evidence of in-group bias with the only significant coefficient on 'Race Match' or 'Gender Match' occurring in Economics where there is a 3.84 mark bonus, equivalent to 0.23 of a s.d. if an exam script had a name indicative of the same ethnicity as the marker. This could further be interpreted as a lack of evidence for taste-based discrimination as one could expect the preferences in taste-based discrimination would be positive towards ones own group but hold 'animus' towards an out-group.

Examining this possible 'animus', columns (5) - (7) add interaction terms with White-Marker, equal to one if the marker was White, as each subject had at least one White marker. These interaction terms have significant effects in the Economics and IT subjects, with White markers grading an exam with a Chinese name 9 marks below a script with a White name in Economics, and 3.7 marks above a White name in IT. This is equivalent to -0.53 and 0.15 of a s.d. respectively. Regressions interacting 'Adopter' with 'White-Marker' were also examined but there were no significant effects found. Columns (5) - (7) show some evidence of taste-based discrimination but is muddied by the different signs in columns (5) and (7), where in IT, White markers prefer Chinese names. Further, there is only one marker in Economics who was White out of the five that graded scripts with names, and two out of five in IT, which means the effects come from a small number of markers. This lessens our ability to make a conclusion about taste-based discrimination.

A.5 Results - Effects of Anonymous Marking

Lastly, we turn to the effects of anonymous marking. Table A3 shows the difference in marks between exams that were marked without any name on the coversheet, and exams that were marked with a name on the coversheet. Therefore, for all columns the omitted group is exams marked anonymously. Column (1) shows the OLS regression coefficients for White, Chinese, and Adopter names where no other control is included, while column (3) is a similar regression but for Male and Female names. Columns (2) and (4) add the controls that are the same as in Table 4, and subject fixed effects, for White, Chinese and Adopter names, and Male and Female names, respectively. Columns (5) and (6), have controls as well as exam script fixed effects. All standard errors are clustered at the marker level.

The results are similar across all columns. There is a strong decrease in marks associated with putting a name on an exam, compared with having an exam marked anonymously. In column (5) the decrease is 9.69 marks from having a White name on the cover sheet compared to no name, and it is an 8.66 mark

Table A3: Anonymous vs. Non-Anonymous

	(1) Mark	(2) Mark	(3) Mark	(4) Mark	(5) Mark	(6) Mark
White	-12.14*	-10.41***			-9.690***	
	(6.072)	(2.963)			(2.690)	
Chinese	-11.23*	-9.405***			-8.656***	
	(6.135)	(2.805)			(2.838)	
Adopter	-10.48*	-8.036***			-9.454***	
	(5.854)	(2.604)			(2.794)	
Male			-10.93*	-9.264***		-9.119***
			(5.951)	(2.556)		(2.751)
Female			-11.67*	-9.333***		-9.430***
			(5.863)	(2.584)		(2.765)
Constant	66.16***	65.10***	66.16***	65.10***	76.76***	76.65***
	(5.532)	(3.564)	(5.528)	(3.534)	(3.173)	(3.073)
Subject Fixed Effects	No	Yes	No	Yes	No	No
Controls	No	Yes	No	Yes	Yes	Yes
Script Fixed Effects	No	No	No	No	Yes	Yes
R squared	0.052	0.18	0.052	0.18	0.88	0.88
Observations	637	636	637	636	636	636
Test of White=Chinese	0.72	0.69			0.18	
Test of White=Adopter	0.41	0.28			0.77	
Test of Adopter=Chinese	0.71	0.47			0.21	
Test of Male=Female			0.47	0.95		0.64

Note: OLS regressions; The dependent variable is the standardised mark an exam received out of 100. Each regression is on a pooled sample of all subjects. The 'Controls' refer to the set of controls used in Table 4. Standard errors are clustered by grader.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

decrease for a Chinese name, a 9.45 decrease for an Adopter name, a 9.12 decrease for a male name, and a 9.43 mark decrease for a female name. All these coefficients are significant at the 1% level. Further, the effect does not differ significantly between name categories. The p-values of tests of equality of coefficients are reported at the bottom of Table A3. The smallest p-value is 0.21, meaning that we cannot reject the null that coefficients do not differ. This suggests that the mark bonus from having an exam marked anonymously benefits all name categories equally.[30]

Now we look at the effectiveness of anonymous marking. When looking at the pooled sample of all subjects in Table 4, exams with Chinese names were found to have a small positive bias compared to White names. However, in column (5) of Table A3, the p-value of the equality of the coefficients of White and Chinese is 0.18 suggesting that anonymous marking has not significantly reduced the bias between these two groups. However, this initial bias found in Table 4 was both small in effect size and only approaching statistical significance. Looking at only the business subject where a larger and more statistically-significant effect was found,[31] anonymous marking significantly reduces the level of bias between White and Chinese names, which is seen in column (2) of Table **??** where the p-value of the equality of the coefficients White and Chinese is 0.085. This infers that anonymous marking is effective for decreasing the largest biases.

B Threshold Effects

B.1 Further Descriptive

Section 3.3 showed histograms which illustrated the change in densities below and above the 50 mark threshold. This is further confirmed in Table B1 which examines the proportion of the distribution of a name groups just below and just above 50. The jump is most pronounced for White and Female names, where only 2% of White names had a mark between 45-49 compared with 13% that had a mark between 50-54. For Females, only 3% get a mark between 45-49, compared to 10% who get a mark between 50-54. Similarly, Table B2 looks at these same distributions from the university's administrative data[32] and shows a similar density of exams with marks just above 50.

[30] Table **??** in the appendix, examines how the effect differs between subject.
[31] See column (2) of Table 5.
[32] Administrative data is examined in detail in Section 4

Table B1: Distributions of marks just below and just above 50

Name Group	45-49	50-54
White	0.02	0.13
Chinese	0.03	0.06
Adopter	0.05	0.08
Male	0.04	0.08
Female	0.03	0.10
Anonymous	0.06	0.07

Note: This table shows the proportion of the distribution of marks of each respective name group that is between 45-49 and between 50-54.

B.2 Anonymous marking

Thresholds effects between exam scripts with names on them were examined in Section 3.3. This section examines them in the context of anonymous exam scripts, to look at the effect of anonymous marking. In Table B3 the dependent variable is equal to 1 if the mark is between 50 and 54, and zero otherwise. In terms of the full sample in column (1) of Table B3 scripts with White names are 7.96 percentage points more likely to receive marks just above 50 than anonymous exams, but the effect is heterogeneous across subjects, where in Economics, White names are 17 percentage points more likely and in IT 9.7 percentage points more likely to have a mark just above 50 than anonymous exams. This is in contrast with Table A3 where on average students are worse off from having a name on an exam. However, in Business, White names are 8.3 percentage points less likely to get a mark just above 50. There are no other significant differences for other ethnic name groups compared to anonymous exams. In terms of gender, in Economics, Females are 16.1 percentage points more likely receive a mark just above 50, while in Business both Males and Females are less likely receive a grade just above 50, while in IT, Males are more likely to receive such a grade.

In terms of the effectiveness of anonymous marking, in a similar way to average effects, anonymous marking reduces the largest biases between name groups. In Table 6, Chinese names were less likely to receive a mark above 50, and in Table B3 anonymous marking reduces this difference, where the test of equality of these coefficients is rejected with a p-value of 0.11. Further, the large positive bias towards females in Economics is also reduced under anonymous marking with a p-value of 0.0078. However, other differences found in column (1) of Table 6, which were smaller effect sizes, are not reduced in Table B3.

Table B2: Distributions of marks just below and just above 50

Name Group	45-49	50-54
White	0.04	0.09
Chinese	0.05	0.11
Adopter	0.04	0.07
Male	0.04	0.10
Female	0.05	0.11
Anonymous	0.05	0.10

Note: Table B1 is redone with administrative data from the university that is described in Section 4. The anonymous distribution is estimated by the distribution in 2018 when all exams were marked anonymously.

Table B3: Threshold Effects (Anon vs. non anon.)

	All Subjects	Economics	Business	IT	All Subjects	Economics	Business	IT
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	P(Above 50)	P(Above 50)	P(Above 50)	P(Above 50)	P(Above 50)	P(Above 50)	P(Above 50)	P(Above 50)
White	0.0796*	0.170***	-0.0832***	0.0976*				
	(0.0407)	(0.0449)	(0.0126)	(0.0406)				
Chinese	0.0199	0.0426	-0.0517	0.00223				
	(0.0370)	(0.0530)	(0.0357)	(0.0245)				
Adopter	0.0287	0.0506	-0.0224	-0.0163				
	(0.0414)	(0.0458)	(0.0259)	(0.0382)				
Male					0.0292	0.0387	-0.0605**	0.0500*
					(0.0332)	(0.0206)	(0.0198)	(0.0199)
Female					0.0577	0.161***	-0.0457**	0.00830
					(0.0390)	(0.0221)	(0.0180)	(0.0188)
Constant	0.0785	0.131	0.0594	0.178	0.0828	0.133	0.0544	0.189
	(0.141)	(0.162)	(0.0405)	(0.268)	(0.133)	(0.134)	(0.0497)	(0.225)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Script Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R squared	0.20	0.25	0.18	0.32	0.20	0.25	0.17	0.29
Observations	636	210	252	174	636	210	252	174
Number of Exam Scripts	101	30	36	35	101	30	36	35
Test of White=Chinese	0.11	0.12	0.53	0.16				
Test of White=Adopter	0.21	0.19	0.025	0.21				
Test of Adopter=Chinese	0.81	0.93	0.64	0.72				
Test of Male=Female					0.29	0.0078	0.70	0.34

Note: The dependent variable equal to one if a mark between 50 and 54 was received. The 'Controls' refer to the set of controls used in Table 4. Standard errors are clustered by grader.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C Robustness Check

C.1 Markers Grading Together

Table C4: Matrix of Correlations for Economics

	Grader 1	Grader 2	Grader 3	Grader 4	Grader 5	Grader 6	Grader 7
Grader 1	1						
Grader 2	0.866***	1					
Grader 3	0.705***	0.694***	1				
Grader 4	0.768***	0.754***	0.659***	1			
Grader 5	0.878***	0.819***	0.669***	0.718***	1		
Grader 6	0.815***	0.768***	0.672***	0.804***	0.832***	1	
Grader 7	0.709***	0.725***	0.573***	0.701***	0.686***	0.710***	1
Observations	30						

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The pairwise correlations between graders for the 30 distinct exam scripts in Economics is reported.

In the experiment, effort was taken to make sure the marker didn't know that there were other marker marking the exams, including emailing each marker individually. Even so, two marker in the Economics subject began grading their bundles of exam scripts together. They noticed an identical exam script in both their piles and went back to the UOS coordinator under the impression some error had been made. The coordinator told them that exams were being double-marked and that they should grade the exams individually. Further, this occurred soon after the two marker received the exam papers so many papers should not have been affected. To examine whether these marker marked the papers together beyond the one identical paper they discovered, the pairwise correlations by distinct exam script are reported in Table C4. The marker were marker 3 and 4, and their r is 0.659, which is among the lowest in the matrix. For marker 3 it is the second lowest correlation with other marker, and for marker 4 it is the lowest correlation with any other marker. This means that it is unlikely that these two marker continued to mark exam scripts together.