國立臺灣大學社會科學院經濟學系
學士班學生論文

Department of Economics
College of Social Sciences
National Taiwan University

Bachelor's Thesis

多元入學造成多媛入學：台灣實證資料

Discrimination Under Non Gender-Blind Tests: Evidence from the Taiwan College Admission

胡翔閎

Sean S.H. Hu

指導教授：王道一 博士

Advisor: Joseph Tao-yi Wang, PhD.

中華民國 111 年 6 月

June, 2022

國立臺灣大學學士班學生論文
口試委員會審定書

多元入學造成多媛入學：台灣實證資料

Discrimination Under Non Gender-Blind Tests: Evidence from the Taiwan College Admission

本論文係胡翔閎君（B06303088）在國立臺灣大學經濟學系完成之學士班學生論文，於民國 111 年 4 月 13 日承下列考試委員審查通過及口試及格，特此證明

口試委員(3 位)：

王道一（簽名）

（指導教授）

黃景沂　　駱明慶

系 主 任 ：＿＿＿＿＿＿＿＿（簽名）

（是否須簽章依各院系規定）

多元入學造成多媛入學：台灣實證資料

胡翔閎．王道一*

中文摘要

我們使用網路上的交叉查榜逾五十萬筆資料，比較 103-110 學年度台灣各大學申請入學通過第一階段篩選的名單，以及第二階段最終錄取的人選，並使用機器學習的方法預測不同人名的性別及名字老氣的程度。以系所固定效應模型所分析的研究結果顯示，第二階段採計愈多「非筆試甄選方式」的科系，錄取名單被機器學習判定為女性的人數比例，較通過第一階段篩選的女性比例增加 5.2%，但在與高中學科相關的基礎科系則較不明顯。相較之下，機器學習判定名字老氣的比例，以及考生所在考區為六都比例，均無類似效果。這表示非匿名的甄選方式 (面試、書審等) 會造成性別篩選的效果，但並未造成其他篩選效果，因此「多元入學」可能造成「多媛入學」，但不一定造成「多錢入學」。我們認為該性別篩選的效果主要來自於男女申請者天生表達能力的差距，然而此一差距在受到足夠訓練以及具備充足背景知識後則大幅縮小，故在於高中教育基礎科目相關的科系中幾乎沒有發現非匿名甄選方式帶來的性別篩選效果。

*作者分別為國立台灣大學經濟學系學士生與國立台灣大學經濟學系特聘教授

Discrimination Under Non Gender-Blind Tests: Evidence from the Taiwan College Admission

Sean S.H. Hu* Joseph Tao-yi Wang†

June 16, 2022

Abstract

One of the most prevailing theories about female underrepresentation in academia, especially STEM, is that recruiting bias against women exists; therefore, tests and assessments in the gender-blind form are often suggested to be applied. Meanwhile, increasing programs implemented interviews reviews on application materials, and other non gender-blind tests for admission after a reform for Taiwan's college admission in 2002. By deeming the evolution in the score weight on non gender-blind tests as a natural experiment, this study examined the effect of the implementation of non gender-blind evaluations on female pupils admission. The empirical result suggests that there is a positive 5.2% female admission change when fully applying non gender-blind assessments, where the average statistic is only 4.0%; moreover, this pro-women effect surges up to 9.6% in those disciplines not directly linked to the core foundation subjects in high school education. We believe this effect is mainly resulted from that women have better expression ability; however, this disparity vanishes upon both genders have the necessary extent of training and understanding in the field. In a nutshell, this research has implications for the debate over which interventions can be a remedy to enhance women's participation in fields where they are outnumbered, and shift this issue to under which conditions some approach is more effective.

JEL codes: I23, J16, J71

Keywords: gender stereotype; leaky pipeline; gender gap; grading bias; higher education

*Department of Economics, National Taiwan University. Email: b06303088@ntu.edu.tw

†Department of Economics, National Taiwan University. Email: josephw@ntu.edu.tw

Contents

List of Figures

List of Tables

1 Introduction

"Try to name 10 female politicians, 10 women CEOs, or simply 10 female scientists from any period in history. Odds are one might run out of names very quickly." A thought experiment adapted from Upson and Friedman (2012) delivers a simple but straightforward message that the female labor force is underrepresented in diverse disciplines and at the same time how often our society neglects this critical issue. Women's equal participation in all aspects of society is one of the most fundamental human rights; however, statistics from various professions only convey frustrating news. Over a century after the first congresswoman was elected, merely around a quarter of parliamentary seats worldwide are currently held by women (Inter-Parliamentary Union, 2021). The business world faces the same obstacle where only around 8% of *Fortune 500* corporations are run by females although it recently reaches a record-breaking female CEOs ratio (Hinchliffe, 2021). One might be even more astonishing how the situation deteriorates in the culinary and movie industry where the Michelin Guide finally has its seventh three-starred women chefs (Marsh, 2021), let alone only one female was ever awarded the best director of the Academy award for over ninety years (Academy of Motion Picture Arts and Sciences, 2021). Among the fields suffering from unbalanced gender participation, academia, especially science, technology, engineering, and mathematics (STEM), is, in fact, the one discussed most extensively (Beede et al., 2011; Leslie, Cimpian, Meyer, & Freeland, 2015; Lundberg & Stearns, 2019), and this problem is often metaphorized as the 'leaky pipeline' (Clark Blickenstaff, 2005). A stark gender disparity will not only dissuade female fresh graduates from pursuing an academic career, but its role model effect can further provoke a profound negative influence on our next generation (Bettinger & Long, 2005). The low female participation rate in most areas of STEM, or to be more specific *male-dominated* fields,[1] is still an urgent issue although the gender gap has been narrowed in recent decades.

Early career aspirations (Sadler, Sonnert, Hazari, & Tai, 2012), personal pref-

[1] Breda and Hillion (2016) note that not all STEM fields are dominated by men.

erences (Ceci & Williams, 2011), and several other theories attempt to clarify the leaky pipeline issue; nonetheless, the most frequent explanation focuses on the prevailing existence of hiring discrimination against females and evidence is documented in both empirical statistics and experimental data. Sheltzer and Smith (2014), for instance, investigated the faculty composition of leading biology laboratories in the United States and noticed that male faculty members, especially the more elite ones, tend to employ fewer female researchers. Moss-Racusin, Dovidio, Brescoll, Graham, and Handelsman (2012) identified male applicants on average were scored significantly as more competent than the (identical background) female applicants by asking faculty from research-intensive institutes to rate students' application materials. Another experiment likewise recorded that subjects are twice more likely to hire a man than a woman candidate in a mathematical task and this bias survives even when the candidates' self-reported performances are disclosed (Reuben, Sapienza, & Zingales, 2014). Related research endorses that hiring discrimination against females prevails; hence, tests and assessments in *gender-blind* form are accordingly suggested to be the proper intervention to preserve gender diversity (Martin & Phillips, 2019; Smith, Handley, Zale, Rushing, & Potvin, 2015). Recent findings, on the contrary, seem to draw a contradicted argument. Ceci, Ginther, Kahn, and Williams (2014), for example, compared male and female faculty tenure-track opportunities and grant funding in the scientific community, and found no compelling gender differences. These related studies expound an unconventional perspective regarding the leaky pipeline debate; nonetheless, a frequent critic is that they might suffer from self-selection bias, i.e., only the most outstanding female Ph.D.'s could survive through the whole process. Williams and Ceci (2015) addressed this issue by sending out fictitious resumes for applying tenure-track assistant professorships, and identified a 2:1 preference in favor of women from their experiment results; however, whether this result has a strong external validity with larger population is still inconclusive. The French research, au contraire, controlled applicants' quality by natural experiments with a larger sample size, yet reaches the same conclusion. Breda and

Ly (2015) analyzed about 3,000 candidates who participated in both interviews and written tests at a top educational institution. Their results show that compared to gender-blind tests (written tests), non gender-blind tests (interview) favor female candidates in male-dominated disciplines. Breda and Hillion (2016) conducted an analysis on roughly 100,000 candidates who participated in the French teaching accreditation exams, and non gender-blind tests are regarded as a counterfactual offer. They pinpoint a 10 percentile rank bias for women in math, physics, or philosophy, where lack of gender diversity. This research suggests that gender discrimination in STEM and academia diminishes; moreover, regarding preventing bias against women, non gender-blind tests should be taken into consideration as they might conversely favor women even in the male-dominated disciplines. Nonetheless, we believe more research is required to verify whether this result holds in different scenarios and populations, and this study aims to fill this gap in the literature. In the meantime, a reform, the College Multiple Entrance program (多元入學方案, hereafter, CME program) introducing non-written tests to the Taiwan college admission, was brought out in 2002 and a boost in female admission rate came along afterward. Before the reform, students predominantly (over 90%) applied to colleges by attending the Joint College Entrance Examination (hereafter, JCEE) consisting of only written tests in several major subjects and they were then allocated to each school by the serial dictatorship mechanism (Satterthwaite & Sonnenschein, 1981) where matching completely depends on their JCEE results. As the original system faded into the background, three main admission channels occurred, namely, "individual application" (個人申請, hereafter, IA), "exam placement" (考試分發), and "star recommendation" (繁星推薦). Among these channels, the first two channels take up more than 85% of the cohort. IA, which consists of interviews, reviews on application materials, and several other non-written tests, increased substantially from 6% of the pupils when first documented in 2001 to almost 60% in 2020 as more and more admission programs[2] tend to recruit students from IA. The overall female

[2]Rather than using the term 'department', we adopt 'admission program' or simply 'program' in this study since each department is allowed to offer multiple admission programs.

student proportions, meanwhile, started a continuous upward trend as non-written tests dominate written tests and become the main evaluation in the Taiwan college admission system; however, the barely increased female applicants rate seems not to account for the single reason for this accession. We hypothesize that it is the increased usage of non-written tests, or to be more specific, the *non gender-blind* assessments that *favors women overall* and leads to the escalation in the female student ratio.

In this study, we analyze Taiwan's CME program data containing over half a million applicants over eight years (2013 to 2020) to verify the hypothesis regarding female pupils growth and explore the effect of applying non/gender-blind tests on the admission bias of the applicants. The *two-stage design* of the CME program which will be later described in Section 2 with *differential usage of non-gender-blind tests* presents a natural experiment our study. Our empirical result records a positive 5.2 female admission percentage point change when completely applying non gender-blind evaluations, where the average statistic is only 4.0; besides, there is no sign of decreasing for this effect in male-dominated disciplines. The overall results suggest that non gender-blind evaluations do not suffocate females' opportunities for entering academia; on the contrary, women gain an advantage in these assessments. Besides, this pro-women effect surges up to 9.65% in those disciplines not directly linked to the core foundation subjects in high school. We suggest that the general effect is due to female pupils' better ability in expression; however, this gap narrows as long as both genders have enough training and understanding in a field.

The remainder of this paper is organized as follows. Section 2 describes our dataset and provides a further introduction to the college admission system in Taiwan. Section 3 demonstrates the identification strategy, whereas the corresponding results are shown in Section 4. We examine the possible mechanisms for this study in Section 5. Last but not least, the paper is concluded in Section 6.

2 Data

2.1 Data Source and Background

This study focuses on analyzing the IA application data since it well portrays Taiwan students' demographic distribution. Our initial dataset is mainly compiled from public information published on a popular college admission results website in Taiwan (交叉查榜網).[3] From this site, information on applicants' Mandarin names, the programs they apply to, and their final placements is collected (Appendix I presents a comprehensive introduction to the main and additional websites and discusses our data reliability). In this study, the research interest predominantly concentrates on the primary channel, IA, in the CME program. As noticed in the foregoing section, IA and 'exam placement' jointly account for over 85% of freshmen each year. Since the latter is held after the former, almost every high school graduates first enter the matching market through IA, and those who are not satisfied with the initial pairing could then reenter the market through the later channel.[4] As a consequence, over 80% of the freshmen at least have attempted to apply through IA every year; thus, the IA pool is a representative sample set of the student population in Taiwan. Not only it is representative, but IA also maintains certain decisive attributes for our research, which will be elaborated in the following paragraph. Table 1 further compares the system before the reform, the current system (IA), and the US university application system.

Table 1: Taiwan/US admission system comparison

	Taiwan past	Individual Application	US admission
Centralized matching	✓	✓	✗
Main test	JCE	GSAT	SAT
Two stage design	✗	✓	✗
Admission office level	-	program	school
Score weight	-	✓	✗

-: Not applicable

[3]We additionally merged the data from another college admission results website (歷屆大學考試分發入學榜單查詢服務) to obtain a better data quality.

[4]Even applicants decide to later reenter the market through 'exam placement', the IA allocation withstands self selection bias since the initial matching results have been documented.

The admission systems in Taiwan and the US have disparities but meanwhile share certain similarities. To begin with, Taiwan's admission system implements centralized matching meaning that matchings between applicants and programs only happen on a unified schedule (Figure 1 presents a detailed timeline) and on the unified platform while schools are allowed to offer early admissions, and applications to each school are usually independent in the latter. The procedure of IA can be divided into two stages where the first phase launches in late January when applicants take the General Scholastic Ability Test (hereafter, GSAT). GSAT can be regarded as an analogy of the SAT Test in the US and it consists of only written tests with mostly multiple-choice questions in major subjects.[5] Applicants can apply to at most *six admission programs* based on their GSAT performances. If they meet the programs' GSAT score requirements, they can advance to the second phase. In the second stage, candidates participate in the evaluations held by each admission program. One should notice that the most fundamental difference between Taiwan and the US college admission system lies in their admission office levels. In IA, each program's admission is determined by its faculty, whereas there is generally a school-level admission office managing the whole institute's admission in the US college system. The evaluations in the second stage can come in miscellaneous forms, e.g., program-specific written tests, interviews, and reviews on application materials. Each admission program is allowed to employ the most suitable (or even multiple) screening approach to evaluate candidates' performance with almost no restrictions, where the only requirement is that the score weights on each evaluation should be predetermined and disclosed beforehand. This makes another difference between the two systems where in the US, most applications will only get a final score based on the whole application package. The declaration of score weights for each evaluation in the second stage of Taiwan's current system enables this study to explore the effect of applying non/gender-blind tests to candidates, and the details will be discussed in the following subsection. Lastly, applicants' scores

[5] Appendix II presents questions from the past GSAT.

(rated independently) for each evaluation will be combinedly calculated to a final outcome based on the score weights set by each admission program. This final score will determine which candidates pass through the second stage and are admitted.

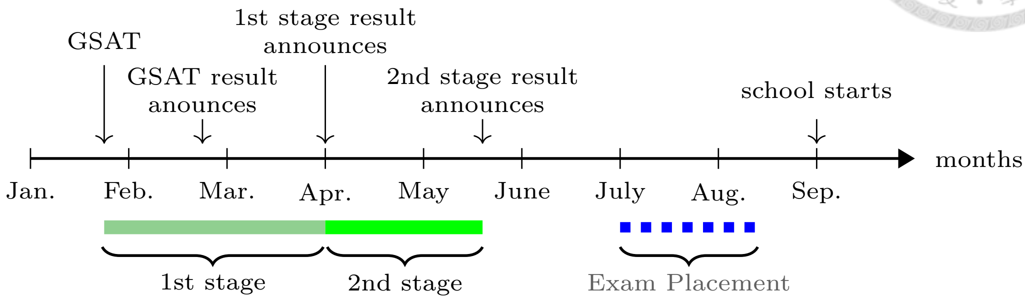

Figure 1: Individual Application schedule

The unique two-stage design in IA further creates a *natural experiment* controlling for applicants' quality in each admission program. That is, students with a GSAT score lower than the program's requirement will be eliminated. Contrarily, since applicants are allowed only to apply to at most six programs, it is more reasonable for a higher score candidate to apply to better-ranked programs rather than being stuck in a much lower one.[6] Consequently, applicants are *sorted* by their GSAT scores in the first stage; therefore, students who pass through the first stage in the same program will have close (or even the same) GSAT scores, i.e., the first stage of IA gives us *good control over candidates quality* at least in terms of written tests performance. Hence, *any shift in applicants' gender distribution after the second stage* is more likely to *singly result from the evaluations in the second stage.*

2.2 Score Weight on Non Gender-Blind tests

A measurement is established for quantifying the extent of gender blindness, i.e., the score weight on non gender-blind tests, for each department in the second stage of IA by the following two steps of calculation. Firstly, each evaluation[7] is labeled with a

[6]It is possible that male pupils are risk-loving and tend to apply to more competitive programs; however, as long as this self-selection is not correlated with programs' usages of non gender-blind tests (which at least cannot be observed in our data), the following analysis will not be affected.

[7]From 2013 to 2020, in terms of names, over 500 different types of evaluations were once implemented.

Non-Gender Blind Score (hereafter, NGB score) by the guided manual classification; that is, an NGB scoring guideline (see Appendix IV for the guideline) is developed then two non-research-related helpers are taught to follow this guideline and rate each type of evaluation an NGB score.[8] The reported scores for every single test are then averaged and (linearly) rescaled to $[0, 1]$. Second, the score weight on non gender-blind tests for each program (hereafter, NGB index) is computed by the weighted summation of the NGB scores for each implemented evaluation based on their corresponding predetermined score weights. To have a better understanding of the calculation of the NGB index, an example is presented below. A program employed evaluation$_a$ and evaluation$_b$ in the second stage, which has a corresponding 0.4 and 0.8 NGB score, and the score weight is 70% and 30%, respectively; therefore, the NGB index for this program is $0.4 \times 0.7 + 0.8 \times 0.3 = 0.52$. Intuitively, if an admission program has a higher NGB index, the applicants' gender are more likely to be revealed in the second stage of IA.

2.3 Features Generating through Machine Learning

One last remaining challenge for our initial dataset is that individuals' gender information is not documented; therefore, machine learning is employed to make gender predictions in this research. Fortunately, similar to the context of English, it is possible to tell an individual's gender from his/her Mandarin name, e.g., a female's name might have a character with the meaning of beauty or simply having the radical "女", which has the meaning of female and woman in Mandarin. Given this desired feature, the predicting model from Liou, Hsiao, Chow, Huang, and Chen (2021) is applied. This model captures the word meaning, word radical, and word pronunciation of an individuals' Mandarin name as the main training features, and random forest, an ensemble supervised machine learning approach, is implemented as the classifier. The corresponding outcomes are in the set of $\{0, 0.1, 0.2, \ldots, 0.9, 1.0\}$ representing the predicted possibility of an individual being female. An individual is

[8] Intuitively, the greater the degree of gender-blindness an evaluation is, the higher this score will be.

labeled as female [male] if this predicted number is greater [smaller] than 0.5 in our study.[9] Since the enforcement of the personal data protection law, the second character in ones' Mandarin names are not disclosed for the most recent two years (2019 and 2020) data.[10] Liou et al. (2021) reported that this approach could reach an overall 93.8% accuracy rate. Besides, there seems to be no clear correlation between the prediction error and individuals' gender; therefore, we believe that the prediction error will only lower the precision (a larger standard error) but not distort the coefficient estimation in our analysis. Moreover, since the following analysis is on program-level, the prediction error for both gender will be balaced out. A better prediction approach could exist; nonetheless, this model has already ensured the predicted gender is dependable for our study.

2.4 Summary Statistics

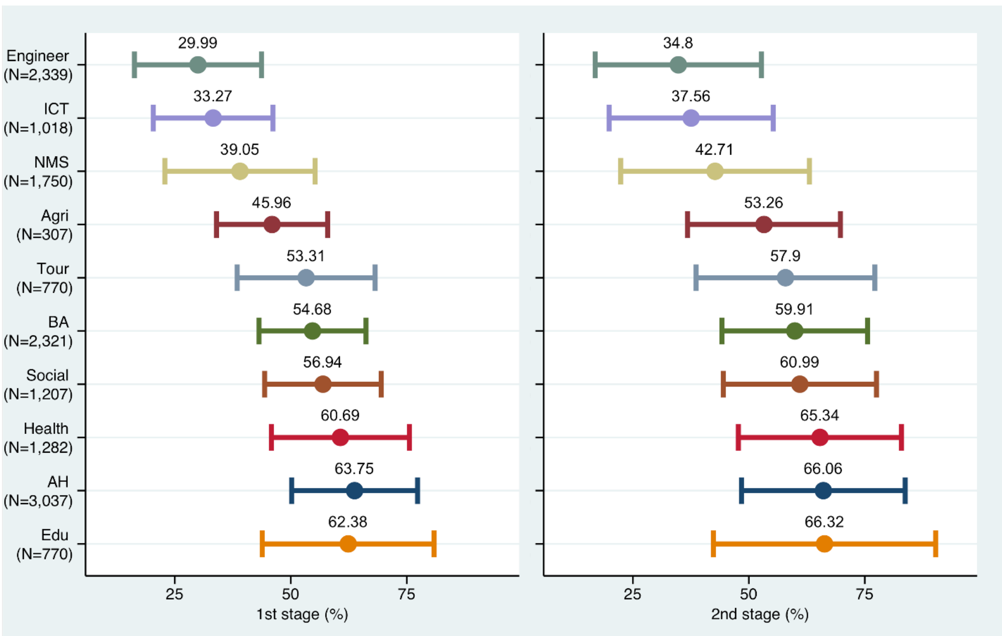

Figure 2: Average female ratio for passing 1st and 2nd stage in major disciplines

Our initial dataset overall contains over half a million (527,581) individuals' ap-

[9]Samples with 0.5 predicted probability are omitted in our main analysis and as a matter of fact, they only accounts a small (<3.8%) share of our data. Detail distribution is presented in Appendix III.

[10]FigureA.4 in Appendix III shows that the prediction is quite consistent across full-information and missing-second-character data.

plication results from most admission programs and universities (2421 programs[11] in 74 universities) in Taiwan between 2013 to 2020. The scale of this data is unprecedentedly considerable compared to the previous studies. Among admission programs, the average female applicant percentage for passing the first stage of IA is 50.19% with 18.87% standard deviation, and it rises up slightly to 54.12% (standard deviation: 21.95%) in the second stage. Figure 2 further summarizes the average female pupil percentage for the first and second stage in major disciplines (see the official website for detailed classification). The abbreviations[12] and number of observations for each field are listed on the vertical axis and the average female pupil distributions are displayed on the horizontal axis. Although there are fewer female applicants in STEM fields (Engineer, ICT, and NMS), an increase in female pupil percentage after the second stage is generally documented in every discipline (ranging from 2.31% in AH to 7.30% in Agri). In addition, the overall average usage of non gender-blind tests (NGB index) calculated from the previous subsections is 0.568 with 0.18 standard deviation. Figure 3 lists the average NGB index in major disciplines. The NMS field has the lowest usage of non gender-blind tests where the highest average NGB index is documented in the Tour field. Most male-dominated fields (Engineer and NMS) have lower usage of non gender-blind tests; nevertheless, only a weak positive correlation ($\rho = 0.06$) is recorded between the female applicants ratio in the second stage of IA and the NGB index. Figure A.5 and A.6 in Appendix V inform the score weights on interview and reviews on application materials, which are the two main components of non gender-blind tests.

[11]Programs are considered as different programs if their names have once changed.

[12]Engineer: Engineering, manufacturing and construction; ICT: Information and communication technologies; NMS: Natural sciences, mathematics and statistics; Agri: Agriculture, forestry, fisheries and veterinary medicine; Tour: Tourism and Catering; BA: Business administration; Social: Social sciences, journalism and library information; Health: Health and social welfare; AH: Arts and humanity; Edu: Education.

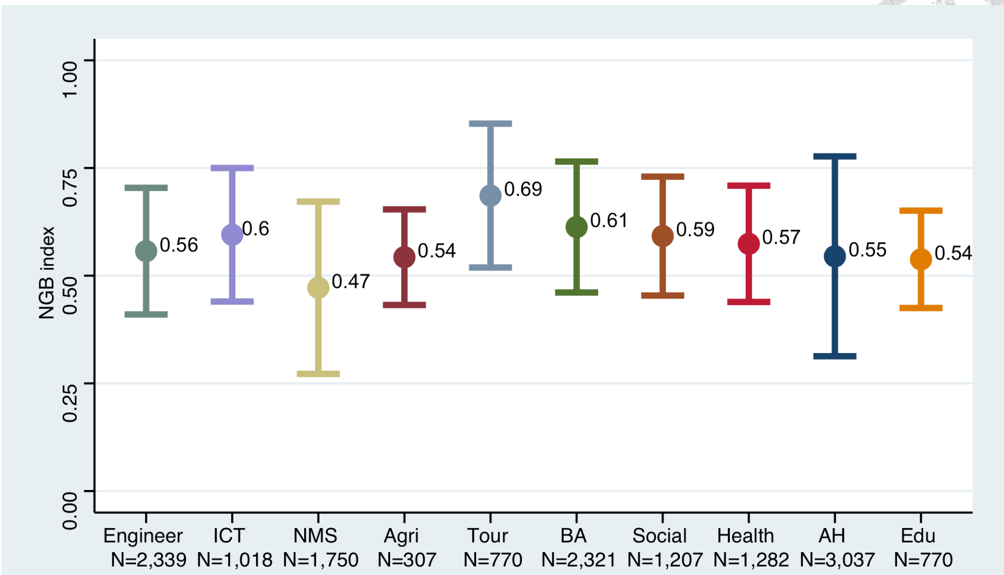

Figure 3: Non gender-blind test usages in major disciplines

3 Methodology

3.1 Regression Model

This study formally analyzes the link between gender distribution evolution in two stages of IA and the individual admission program's differential usage of non-gender-blind tests using regression specifications of the following form:

$$\Delta \text{ female}\%_{ij} = \beta \text{NGB index}_{ij} + X_{ij}\Gamma' + \varepsilon_{ij}, \tag{1}$$

where i and j index the admission program and year, respectively. Δ female% is the female ratio difference between the candidates who passed the second stage and the first stage. If Δ female% is positive, the female ratio increases after the second stage of IA, vice versa. NGB index is the score weight on non gender-blind test obtained from the previous calculation. The changes in NGB index for each program, to our best knowledge, are not correlated to the intention to recruit more specific gender students; in fact, changes in NGB index mostly occur because of the demand to better differentiate applicants in the second stage of IA. In some specifications, the admission program and year fixed-effects are controlled to guarantee our results are not simply the preference differences between each admission program.

4 Empirical Results

4.1 Main Results

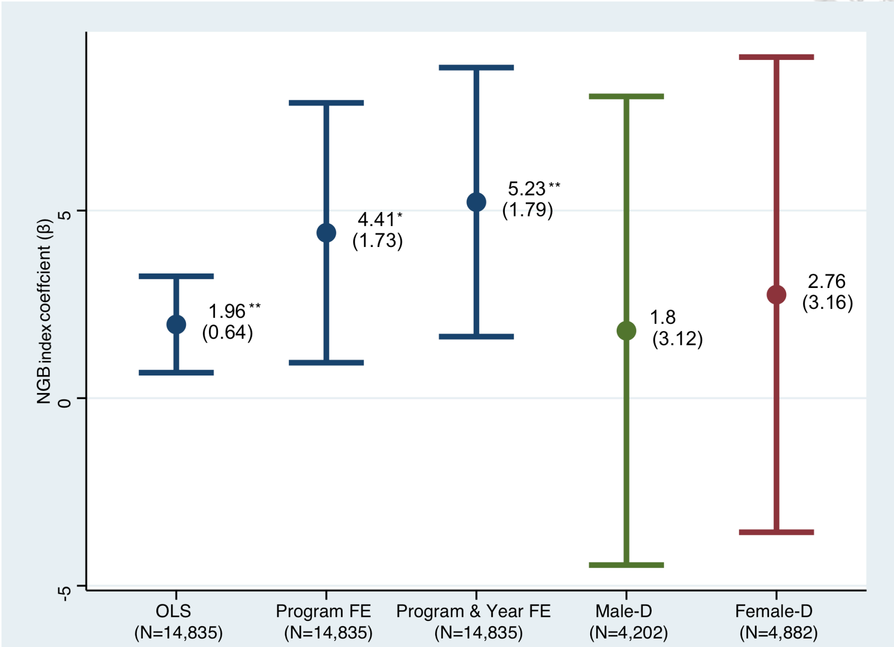

Figure 4: Female evaluation advantage under non gender-blind tests. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

Figure 4 documents the coefficient for NGB index (β) from equation (1) under different specifications, i.e., the average female pupil increment rate after the second stage when fully applying non gender-blind tests compared to gender-blind ones. The first column constitutes only the NGB index but no other controls, and a very weak correlation between the difference in female candidates rate (Δ female$\%_{ij}$) and the NGB index is observed; however, with the inclusion of the individual program fixed-effect in the remaining models, a positive relation occurs. As demonstrated in the third column, when both individual program fixed-effect and year fixed-effect are controlled, if an admission program completely applies non gender-blind tests in the second stage, there will be a 5.23 percentage point change increase in the admitted female student ratio compared to gender-blind ones.The fourth [fifth] column applies the same specification in the third column and examines the non gender-blind effect in male-dominated [female-dominated].[13] The estimation informs that there is only

[13] A program is classified as male-dominated [female-dominated] if and only if the average female

a slight drop in the effect size for gender-specific fields compared to the general field, and the estimation has no significant difference (at least economically) between both subsamples. The number of observations for each specification is listed in each parenthesis under each column.

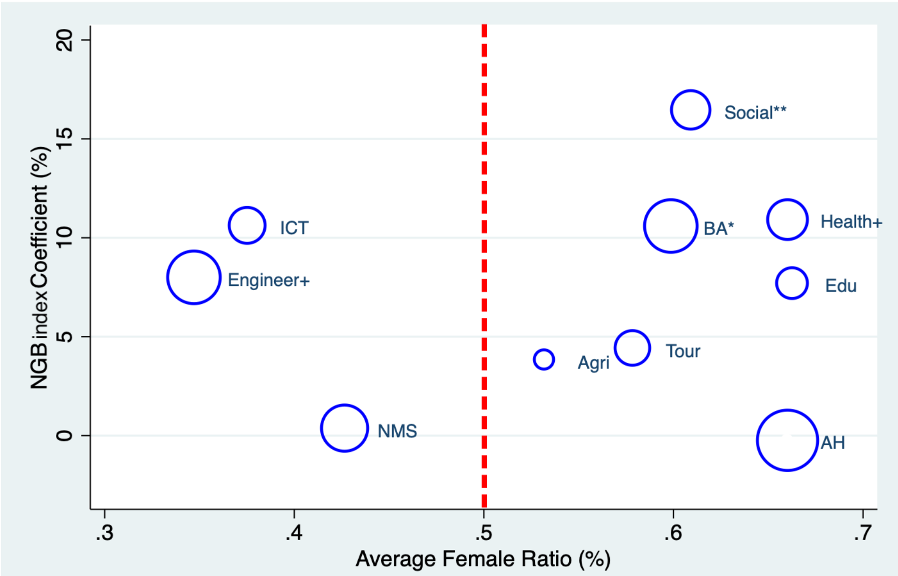

Figure 5: Female evaluation advantage under non gender-blind tests in major disciplines and disciplines' extent of male-domination. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05, +: p<0.1)

Following the same specification in column (3) of Figure 4, we explore the effects of non gender-blind tests in major disciplines (same discipline classification with Figure 2) and the results are reported on the y-axis of Figure 5. The size of each circle indicates the number of program samples included in the corresponding discipline, and the extent to which each estimated coefficient is different from 0 is reported by the star sign following the label. Lastly, the disciplines' extent of male-domination (x-axis) is measured by the average female student ratio after the second stage of IA. From Figure 5, the pro-women favor of non gender-blind evaluation is observed in almost every field (mostly located between 5% to 15%), and the effect size can even climb up to over 15% in the field of Social sciences, journalism and library information (Social). By the division of the middle red dash line, there is no clear evidence that women are favored/disadvantaged in more male-orientated nor

applicants ratio for the the first stage is greater [smaller] than 60% [40%].

female-orientated areas, which corresponds to the estimation in column (4) and (5) from Figure 4. Another interesting finding is that non gender-blind evaluation effect vanishes in both Natural sciences, mathematics and statistics (NMS) and Arts and humanities (AH). One feature both disciplines share is that they are both directly connected with the core foundation subjects in high school. In Taiwan, mathematics, natural science (physics, chemistry, biology), Chinese, English, and social studies (geography, history, civics) are regarded as the core foundation subjects in high school education and are as well as the subjects in GSAT. Under the official classification, programs studying the above subjects are mainly in either NMS or AH.[14] This discovery possibly offer an opportunity to shed light on the mechanism behind this pro-women favor of non gender-blind evaluation, and further discussion is presented in the following section.

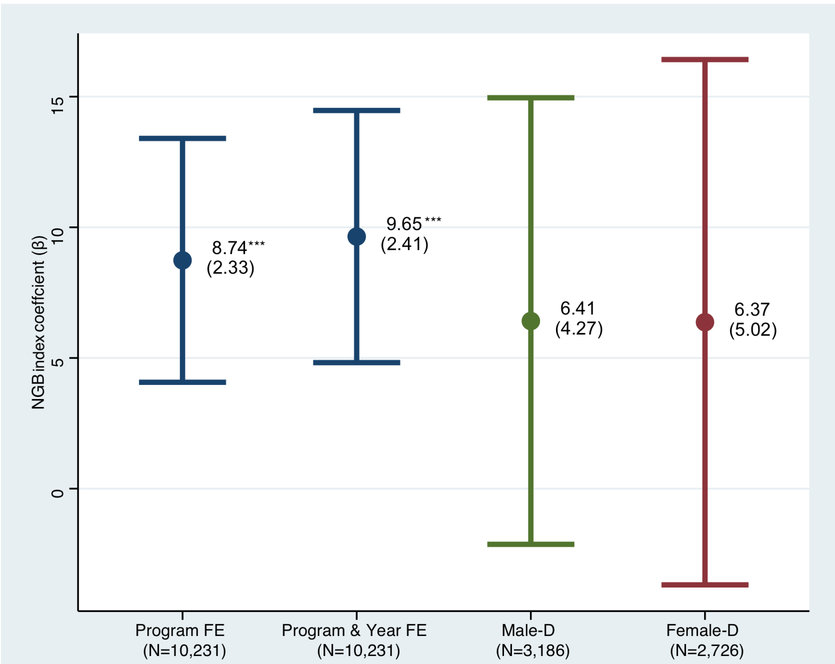

Figure 6: Female evaluation advantage under non gender-blind tests w/o NMS and AH. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

Figure 6 adapts the same specifications with Figure 4 yet on the sample set with the exclusion of NMS and AH data. The non gender-blind effect escalates to over 8.74% in the overall data when only the individual program fixed-effect is

[14] Civics is classified in Social sciences, and Geography is sometimes not classified in NMS.

controlled, and to 9.65% with the further inclusion of the year fixed-effect, where the average change between the first and the second stage of IA is only 4.64% under this subsample. Column (3) and (4) from Figure 6 as well record a strong correlation between admitted female pupils and non gender-blind assessment usage in both male-dominated and female-dominated admission programs. In short, a statistically significant positive correlation between the NGB index and the difference in female candidates in the two stages of IA is observed in general data; nonetheless, the non gender-blind effect tends to decline in disciplines closely connected to high school core foundation subjects.

4.2 Interpretation of Magnitudes

The foregoing discussion focused on the analysis of the effect of the usage of non gender-blind tests on female pupils admission. In this section, we ask whether the above effect sizes are large or small in absolute terms. The average difference between the female rate for students passing the second stage and that in the first stage is 4.08% while the overall effect size captured in the preceding section[15] is 5.23%. This comparison shows that when a program fully applies non gender-blind tests in the second stage, it can recruit approximately 1.28 times more female students. This phenomenon grows even larger for the sample set with the exclusion of NMS and AH disciplines. The 8.74% effect size is 1.88 times greater than the average number (4.64%) in these disciplines. In the following, we provide two examples to further illustrate how much the non gender-blind usage can influence the female admission. Firstly, in 2019, the department of Civil Engineering in National Taiwan University (hearafter, NTU) applied fully gender-blind test (written tests) in the second stage of IA (NGB index = 0). A year after that, the department accommodated its admission strategy with the inclusion of 30% of non gender-blind evaluation (NGB index becomes 0.3). Based on the above analysis and simple linear model prediction, the predicted difference between two years' female admission rate should be 0.3 ×

[15]The third column in 4.

$5.23\% = 1.584\%$ which almost coincides with the actual difference (1.455%). On the contrary, in the very same year the department of Mechanical Engineering in NTU reduces the usages of non gender-blind evaluation with an accordingly 0.2 reduction in NGB index. By again applying the above linear model, we expect it should suffer a $0.2 \times 5.23 = 1.046\%$ decrease in female admission ratio. In hindsight, the department indeed experienced a 2.858 percentage loss in its female admission ratio. The above two examples illustrate that how the adjustment in the usage of non gender-blind evaluation could affect departments' admission. In addition, by only implementing naive linear model, our analysis could already produce reliable estimation based on departments' past NGB index.

4.3 Further Studies

Non gender-blind tests could be further categorized into two main groups of assessment, interview and review on application materials, and this division leads to an intriguing issue, that is, which test plays a more important role in the non gender-blind effect documented in Figure 4 and 6. A common intuition makes a naive assumption that this correlation could mainly result from that women's better oral expression ability; therefore, they could easily stand out in interviews. To verify this hypothesis, the models in Figure A.7 in Appendix VI follow the same specification in Figure 4, but only the score weight on non gender-blind tests (NGB index) is now separated into the score weights on interviews and reviews on application materials. The blue dots represent the effect size of reviews on application materials while the red dots represent the stats for interview. The result from Figure A.7 rejects the above hypothesis since there is as well an equal-size effect coming from reviews on application materials. A similar analysis on the dataset with the exclusion of NMS and AH is also presented in Appendix VI. We again observe the equal effect sizes for both evaluations are reached.

The preceding paragraphs have suggested a positive correlation between non gender-blind test usage (NGB index) and admitted female student rate. Below we proceed by examining the non gender-blind test effect in two subsamples. First and foremost, programs in each tier could have varied admission strategies. On the one hand, the admission for top-ranked programs, for instance, is highly competitive, allowing their examiners more flexibility in focusing on singling out suitable and outstanding candidates; on the other hand, since matching markets in lower-ranked programs are relatively thin and most applicants only regard these programs as a backup option, the priority for these programs will mostly address on identifying those who will actually stay if the offer is granted. Given this diverse strategy assumption, it will be, therefore, intriguing to verify whether the non gender-blind test effect will alter in different tiers. Breda and Hillion (2016), as a matter of fact, identify the pro-women non gender-blind test bias takes effect in higher-level evaluation (professorial and high school teaching accreditation exams) but wears off in medium level (secondary school). In Appendix VII, the identical specifications from Figure 4 are undergone on programs in three separated tiers (see Appendix VII for categorization details). Our analysis suggests that the non gender-blind tests tend to have a universal effect in all tiers and there is no significant difference (at least economically) between their effect sizes, which in a sense seems to diverge from Breda and Hillion (2016)'s finding. Therefore, we believe the pro-women effect found in this research is consistent in different ranked programs. The second subsample analysis is concentrated on the programs' size. We only focus on the change in female applicant ratio, whereas the program size has not been taken into consideration in the above analysis; nonetheless, this might pose a threat that the empirical results could be distorted by programs with fewer candidates.[16] The same specifications from Figure 4 and 6 are undergone and weighted by programs' sizes (analytic weights) in Appendix VIII to clarify the above issue. In Figure A.10 and A.11, the effect size for non gender-blind test in overall data indeed decreases; nevertheless, a positive relationship still exists and the coefficients are statistically significance different from zero.

[16]For instance, admitting one more female applicant increases the gender ratio in a 10-student size program by 10%, yet only 2% in a 50-student program.

We have studied the effect of non (gender-)blind tests on female admission and do capture a positive correlation. One can imagine that it is possible that examiners could infer candidates' living region from application materials or how rich his/her family is by their outfit in a non blind test; thus, whether examiners will accordingly adjust their preferences based on these two non-ability-related factors could be another crucial study. In the very last paragraph of this section, we perform two additional analyses on whether applying non blind tests in the second stage of IA will influence the admitted students' demographic characteristics on urban-rural distribution and socioeconomic status. First of all, for the urban-rural distribution, the locations where each pupil takes the GSAT are documented by the same admission results website, and this information is implemented as a proxy to classify whether an individual is living in the urban area. Students are assigned to a GSAT location in the same city where their high school locates. Besides, according to the data from the Ministry of Education, in 2010, over three-quarters of the high school cohort registered in a hometown high school; therefore, the exam location could be a good approximation of the place where an individual lives. As for socioeconomic status, we apply another machine-learning-based prediction model established in Liou et al. (2021) for predicting individuals' age based on their Mandarin name. Levitt and Dubner (2014) discuss that there is a clear pattern that names chosen by wealthy families will later appear in average or poor family children; therefore, names could provide a signal of individual socioeconomic status. Given the applicants in the same year are mostly fresh high school graduates and at the same age, we hold the assumption that an applicant who has an older name-predicted age is more likely to have lower socioeconomic status since s/he are more likely to stand at the end of this name-passing channel, vice versa. Figure A.14 in Appendix IX exhibits no correlation between the usage of non-blind tests and students' urban-rural distribution nor socioeconomic status, meaning that the non blind test effect might only exist in the aspect of gender distribution but not other factors.[17]

[17]Further details are discussed in Appendix IX.

5 Discussion

The results from the previous section demonstrated a clear relationship between the implementation of non gender-blind evaluations and female admission. In contrast to to most previous studies, our empirical results suggest that non gender-blind tests do not suffocate women's opportunities for academic participation yet even conversely favor them. In this section, we focus on discussing the possible mechanisms behind our findings. Below we propose three possible mechanisms to explain the above results and examine whether each of them is consistent with our analysis respectively: (i) there might exist self-selection for applicants in IA. However, candidates are sorted by the first stage control based on students' GSAT scores. Moreover, the analysis on different school tiers (Appendix VII) indicates that even in lower-tier schools where the selection system is relatively less effective, the non gender-blind test effect prevails nevertheless. We can, hence, reject applicants' self-selection being one of the mechanisms behind our findings. (ii) examiners (faculty) have a taste in favor of female applicants. However, if this argument is true, it further imply examiners in core disciplines should have distinctive tastes diverging from those of non-core discipline faculty; hence, there is no pro-women effect in core disciplines. Genuinely speaking, more delicate individual level score information in the second stage assessments is required to further verify this argument; meanwhile, there is as well no compelling rationale to conclude that taste disparity exists. One should notice that the current literature has only established a clear pattern between gender bias/discrimination and gender ratio in specific fields (e.g. STEM), yet the relationship between gender bias/discrimination and other factors remains unclear. (iii) applicants' performances in the second-stage may still vary even after the first stage selection. The first-stage selection is only based on students' GSAT records, i.e., only the ability that can be examined through written tests (mostly with standard answers) will be controlled. In the second stage, each program will design its program-specific assessments to recruit the most suitable candidates. This is, in fact, the main spirit of IA where the whole system is designed to single out those pupils who are potentially well-fitted to a specific program rather than those who only have higher test scores. Because of earlier puberty, teenage females usually are regarded as having a better ability to meaningfully express their thoughts and ideas through appropriate syntactic and semantic, whether by oral (interviews) or writing (preparing application materials) compared to the same age males; therefore, there is no wonder a strong positive correlation between the extent of application on non gender-blind tests and female admitted ratio is captured in almost every field. One should, nevertheless, notice that this relationship evaporates in those disciplines closely connected to the core foundation subjects in high school education (hereafter, core disciplines). The most fundamental difference between programs in core disciplines and those not in them is that throughout the three years of high school, education is mostly focused on expanding students' understanding of the subjects in core disciplines, whereas the related knowledge of other fields remained untouched. That is to say, applicants in core disciplines will usually be equipped with a greater extent of related prior knowledge in the second stage of IA compared to those in other disciplines will. Under this discrepancy, no clear relationship between the implementation of non gender-blind tests and female admission is observed in the core disciplines. We believe that the expression advantage to women fades when both genders have enough level of understanding in the field; hence, the non gender-blind effect diminishes in core disciplines. To be more specific, the below Figure 7 lists the minimum required credits for each subject in Taiwan's high school education. We can see a clear pattern that when the minimum required credits is greater than or equal to four units (orange-colored bar) for a subject, the non gender-blind tests effect in a related discipline (NMS and AH) is not significant different from zero;[18] on the contrary, when a subject have a minimum required credits that are smaller than four units (blue-colored), the non gender-blind effect in the related disciplines appears. Notice that the only exception lies in Civics. Civics has a total eight minimum required credits; however, it actually includes four subjects (Sociology,

[18]Although Art belongs to AH, but usually only high school graduates from art schools will attend those programs studying art.

Political science, Law, and Economics) taught across four semesters, in which one of them is covered in each semester. In other words, since Civics covers a broad range of subjects in Social science, students can only gain a basic level of knowledge in each individual subject; thus, it turns out that non gender-blind effect is still documented in these programs.

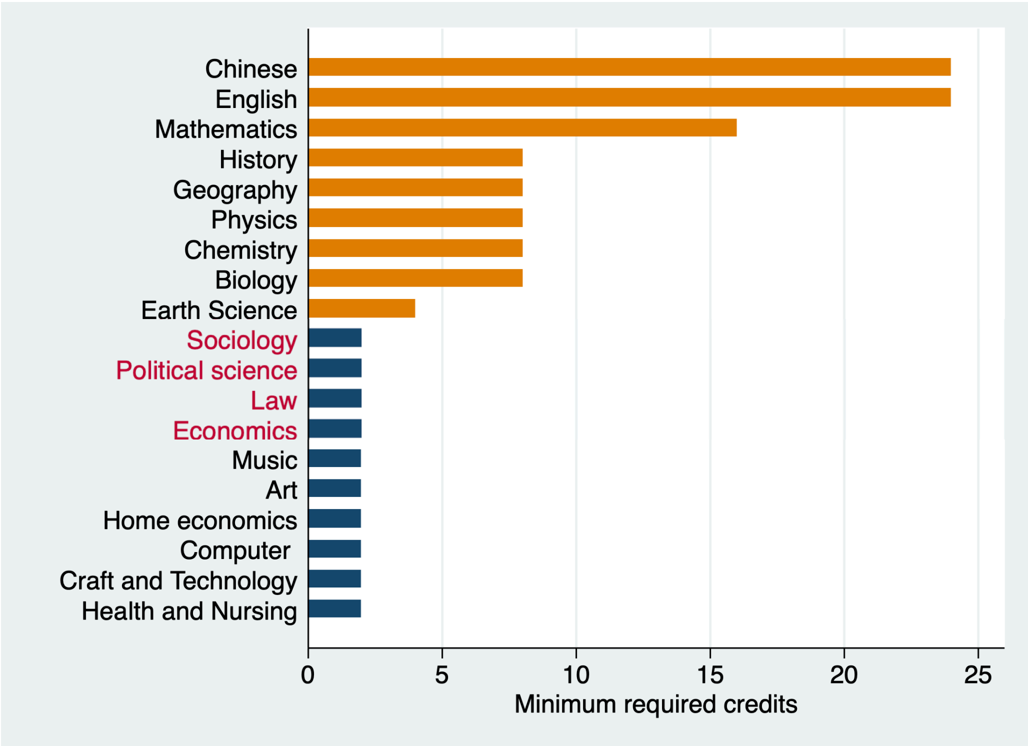

Figure 7: Minimum required credits in each subject

6 Conclusion

This study investigates the effect of applying non/gender-blind tests on female admission in academia by examining college admission data from Taiwan. Based on the two-stage design in IA, we discover that there is a 5.23 female admission percentage point increase when non gender-blind tests are implemented, where the average statistic is only 4.08.; besides, there is no sign of a significant decrease in specific gender-dominated disciplines. The overall results suggest that non gender-blind evaluations do not suffocate females' opportunities for entering academia as opposed to most previous studies stated; on the contrary, women can even be favored in these assessments. Moreover, this pro-women effect surges up to 9.65% in those disciplines with non-direct connection with the core foundation subjects in high school. We believe the non gender blind test effect captured in our research mainly arises from female pupils' expression ability advantage; however, this gap narrows whenever both genders have enough training and knowledge in the field. In a nutshell, this study presents a meaningful insight into related policies: Employing non gender-blind tests can be an efficient approach to increase the representation of women in general academic fields, and the leaky pipeline debate should shift from which intervention is appropriate to under which conditions some approach is more effective.

References

Academy of Motion Picture Arts and Sciences. (2021). *The official academy awards database.*

Beede, D. N., Julian, T. A., Langdon, D., McKittrick, G., Khan, B., & Doms, M. E. (2011). Women in stem: A gender gap to innovation. *Economics and Statistics Administration Issue Brief*(04-11).

Bettinger, E. P., & Long, B. T. (2005). Do faculty serve as role models? the impact of instructor gender on female students. *American Economic Review*, *95*(2), 152–157.

Breda, T., & Hillion, M. (2016). Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in france. *Science*, *353*(6298), 474–478.

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.

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in academic science: A changing landscape. *Psychological science in the public interest*, *15*(3), 75–141.

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, *108*(8), 3157–3162.

Clark Blickenstaff, J. (2005). Women and science careers: leaky pipeline or gender filter? *Gender and education*, *17*(4), 369–386.

Hinchliffe, E. (2021). The female ceos on this year's fortune 500 just broke three all-time records. *Fortune.com.*

Inter-Parliamentary Union. (2021). *Global and regional averages of women in national parliaments.*

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. *Science*, *347*(6219), 262–265.

Levitt, S. D., & Dubner, S. J. (2014). *Freakonomics*. B DE BOOKS.

Liou, J.-S., Hsiao, C.-Y., Chow, L.-Y., Huang, Y.-H., & Chen, Y.-S. (2021). Age estimates from name characters. *Applied Sciences*, *11*(20).

Lundberg, S., & Stearns, J. (2019). Women in economics: Stalled progress. *Journal of Economic Perspectives*, *33*(1), 3–22.

Marsh, S. (2021). Clare smyth and hélène darroze celebrate three michelin stars. *The Guardian*.

Martin, A. E., & Phillips, K. W. (2019). Blind to bias: The benefits of gender-blindness for stem stereotyping. *Journal of Experimental Social Psychology*, *82*, 294–306.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty' s subtle gender biases favor male students. *Proceedings of the national academy of sciences*, *109*(41), 16474–16479.

Reuben, E., Sapienza, P., & Zingales, L. (2014). How stereotypes impair women' s careers in science. *Proceedings of the National Academy of Sciences*, *111*(12), 4403–4408.

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of stem career interest in high school: A gender study. *Science education*, *96*(3), 411–427.

Satterthwaite, M. A., & Sonnenschein, H. (1981). Strategy-proof allocation mechanisms at differentiable points. *The Review of Economic Studies*, *48*(4), 587–597.

Sheltzer, J. M., & Smith, J. C. (2014). Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences*, *111*(28), 10107–10112.

Smith, J. L., Handley, I. M., Zale, A. V., Rushing, S., & Potvin, M. A. (2015). Now hiring! empirically testing a three-step intervention to increase faculty gender diversity in stem. *BioScience*, *65*(11), 1084–1087.

Upson, S., & Friedman, L. F. (2012). Where are all the female geniuses? *Scientific American Mind*, *23*(5), 63–65.

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

Appendix

Appendix I: 交叉查榜網

Every year after the second stage results of IA are released by each program and school, this website will collect all of this information. This information will then be organized in the setup like the below graph presented. For each individual, the website lists the program s/he chose to apply and the results for each program. Applicants can acquire the above information for the applicants who are ranked higher in the same program and estimate whether s/he will end up choosing another (usually better ranked) program, which increases the possibility that this applicant could be admitted. The information for a few candidates' Mandarin names is incomplete (e.g., the second character is missing); therefore, we merged the admission data from another college admission result website (see 歷屆大學考試分發入學榜單查詢服務 for details)[19] by candidates' ID numbers to obtain a better data quality. For the data reliability, we compared both datasets from two individual websites and the comparison showed high data overlapping; therefore, this offers evidence of the reliability of the data from the original website.

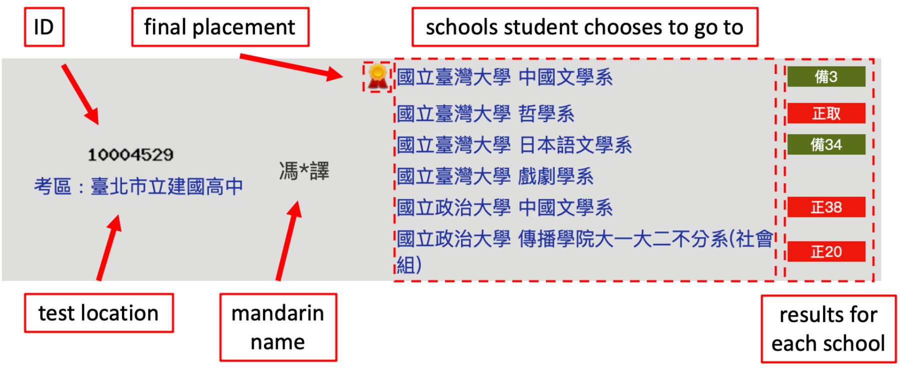

Figure A.1: Screenshot from 交叉查榜網

[19]The reason why this data is not applied in our main analysis is that it contains only the information for final placements but not candidates who passed the second-stage. Final placements are determined by applicants but not examiners; hence, it suffers severe self-selection.

Appendix II: Past questions from GSAT

二、綜合測驗（占15分）

說明：第16題至第30題，每題一個空格，請依文意選出最適當的一個選項，並畫記在答案卡之「選擇題答案區」。各題答對者，得1分；答錯、未作答或畫記多於一個選項者，該題以零分計算。

第 16 至 20 題為題組

You begin to notice a bit of pain on your eyelid each time you blink. You __16__ the mirror to find a tiny red spot on the base of your lower lashes. These __17__ are probably the beginning of an eye stye.

An eye stye is a small bump, resembling a pimple, that develops when an oil gland at the edge of an eyelid becomes infected by bacteria. These bacteria are found in the nose and are easily __18__ to the eye when you rub your nose, then your eye. Pus will build up in the center of the stye, causing a yellowish spot. Usually a stye is accompanied by a swollen eye.

__19__ a stye can look unpleasant at times, it is usually harmless and doesn’t cause vision problems. Most styes heal on their own within a few days. You might speed up healing time by gently pressing a warm washcloth __20__ your eyelid for 10 minutes, 3 or 4 times a day. Make sure you don’t squeeze or pop a stye like you would a pimple. Doing so may cause a severe eye infection.

16. (A) check out (B) look into (C) watch over (D) see through
17. (A) incidents (B) measures (C) symptoms (D) explanations
18. (A) attracted (B) contributed (C) exposed (D) transferred
19. (A) As (B) If (C) Unless (D) Although
20. (A) against (B) among (C) about (D) after

第 21 至 25 題為題組

Shoes are hugely important for protecting our feet, especially in places like Africa, where healthcare provision is limited. Unfortunately, shoes are not always readily available for people living in poverty, __21__ shoes that are the right size. Almost as soon as a child receives shoes to wear, he/she is likely to have grown out of them. Then the child has to __22__ with shoes that are too small. *The Shoe That Grows*, created by a charity called Because International, changes all this. It allows children to __23__ their shoes’ size as their feet grow.

The innovative footwear resembles a common sandal and is made of leather straps and rubber soles, a material similar to that used in tires. They come __24__ two sizes, and can expand in three places. The straps on the heel and toe control the length of the shoe, __25__ the two on either side allow for different widths. With this special design, the shoes can “grow” up to five sizes and last for at least five years.

21. (A) except for (B) provided with (C) far from (D) let alone
22. (A) get done (B) get lost (C) make do (D) make believe
23. (A) adjust (B) explore (C) insert (D) overlook
24. (A) by (B) in (C) from (D) down
25. (A) whether (B) while (C) with (D) for

Figure A.2: GSAT past questions

Appendix III: Female prediction distribution

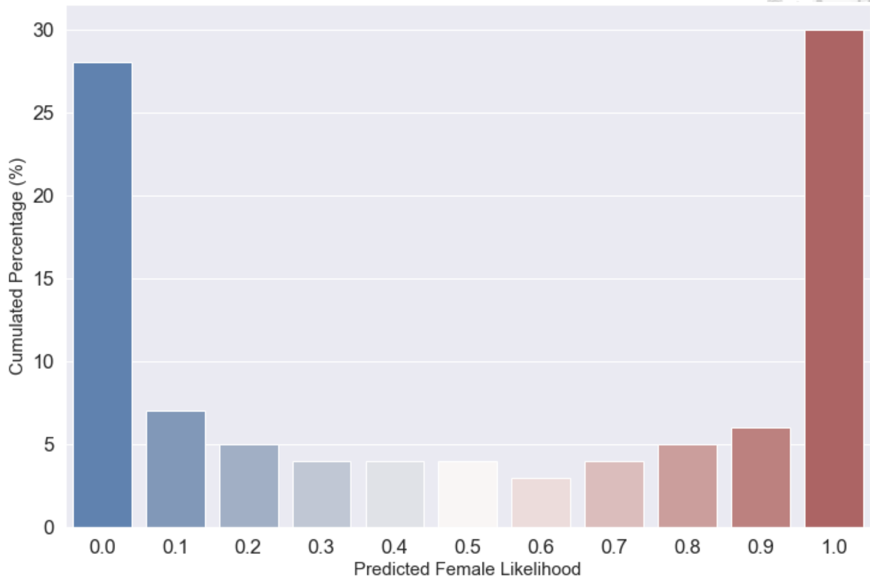

Figure A.3: Prediction results distribution

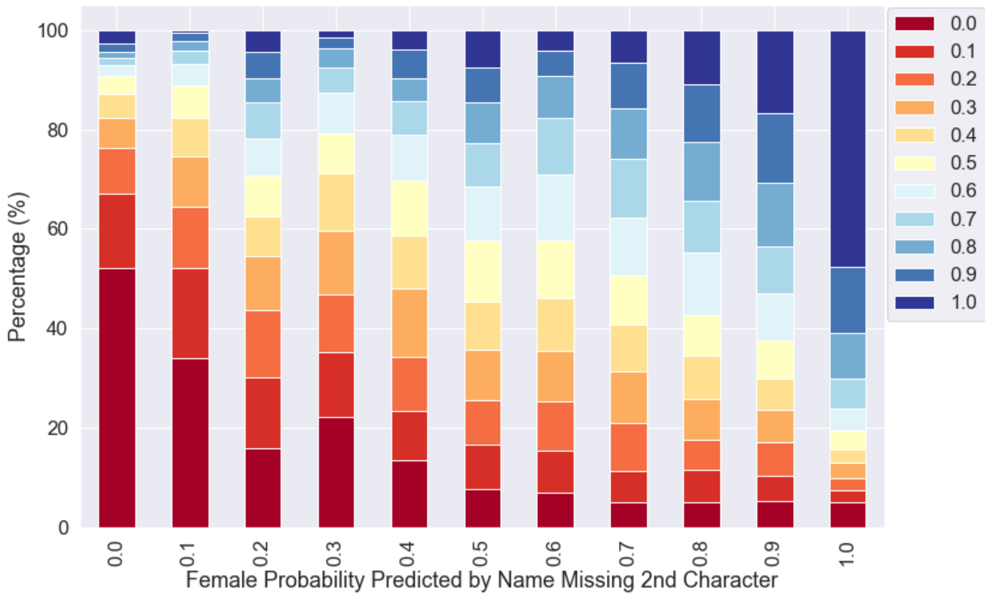

Figure A.4: Comparison between full-information or missing-second-character data

Figure A.4 compares the prediction outcomes of the two sets of data. The first set of data is composed of full information; the second one is composed of only missing-second-character data. The x-axis represents the female probability predicted by names missing the second characters. The y-axis represents the proportion of the samples predicted to be a specific value when using complete data. This graph shows that the prediction performance difference between full-information or missing-second-character names is not that significant at least in terms of gender classification.

Appendix IV: NGB score guideline

[Original Mandarin version]

1. 性別盲定義：如果考官無法藉由該種考試方式得知考生性別，則該考試方式即為性別盲，反之，如果考官可以得知考生性別，則該考試方式即為非性別盲。

2. 給分：

   - 非常肯定該考試為性別盲：該考試格位填入 0
   - 無法肯定、但認為該考試應較可能為性別盲：該考試格位填入 1
   - 該考試為性別盲、非性別盲可能程度幾乎相同：該考試格位填入 2
   - 無法肯定、但認為該考試應較可能為非性別盲：該考試格位填入 3
   - 非常肯定該考試為非性別盲：該考試格位填入 4

3. 備註：
   如不確定該考試的進行方式，可以自行上網搜尋相關資料確認。也可以直接查訊使用該考試方式學系的相關規定。例如：根據師範大學網站介紹「APCS」為「大學程式設計先修檢測」，藉由舉辦具公信力之「程式設計檢測」，讓具備程式設計能力之高中職學生，能夠檢驗學習成果，並供作大學選才的參考依據。

[Translated English version]

1. Definition of gender blindness:
   If examiners cannot tell applicants' gender by a test method, then this examination method is gender blind. On the contrary, if the examiner can tell applicants' gender, the examination method is of non-gender blind.

2. Gender Blind Score:

   - Very sure that the test is gender-blind: fill in 0 for the test.
   - Uncertain, but more likely to be gender blind: fill in 1 for the test.
   - Uncertain, the possibility for being gender-blind and non gender-blind is almost the same: fill in 2 for the test.
   - Uncertain, but more likely to be non gender blind: fill in 3 for the test.
   - Very sure that the test is non gender-blind: fill in 4 for the test.

3. Remarks:
   If you are not sure about how the exam will be conducted, you are allowed to search for relevant information on the Internet. For example, according to the Internet, "APCS" is the "Preliminary Test of Programming for Universities". By organizing a credible "programming test", high school vocational students with programming ability can test their learning results and provide them for use. Reference basis for university selection.

Appendix V: Score weights on interviews and reviews on application materials

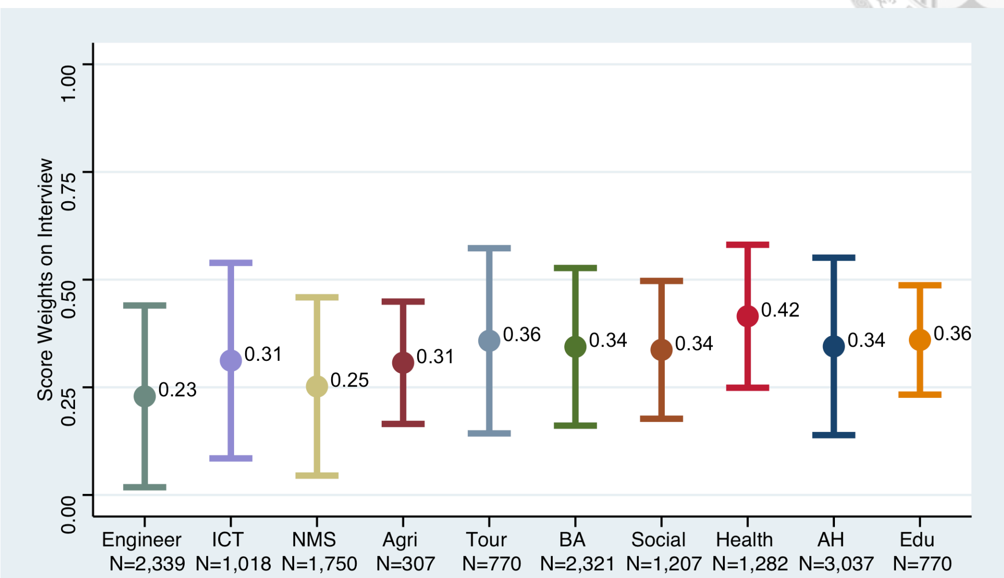

Figure A.5: Score weights on interviews in major disciplines

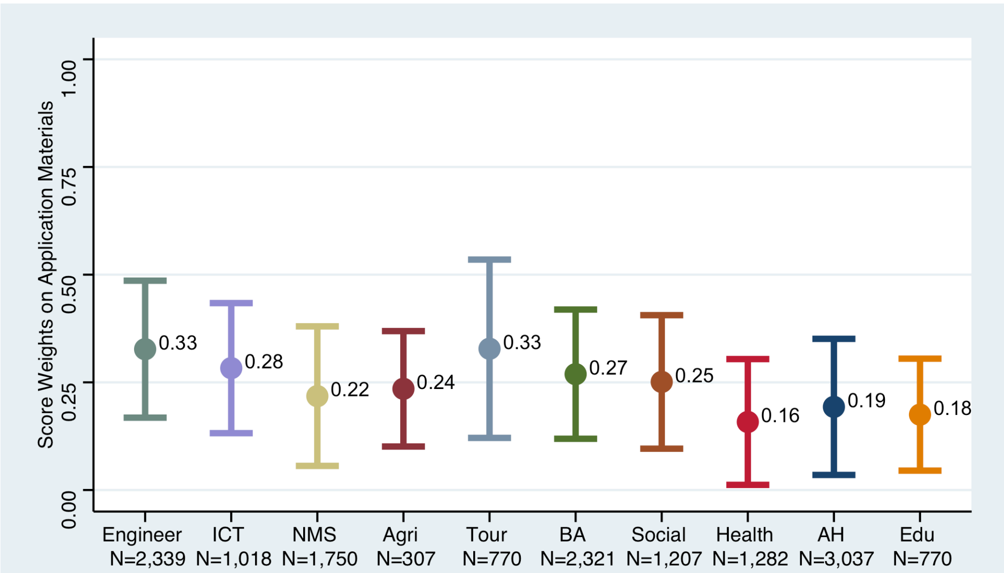

Figure A.6: Score weights on reviews on application materials in major disciplines

Appendix VI: Analysis on divided NGB index

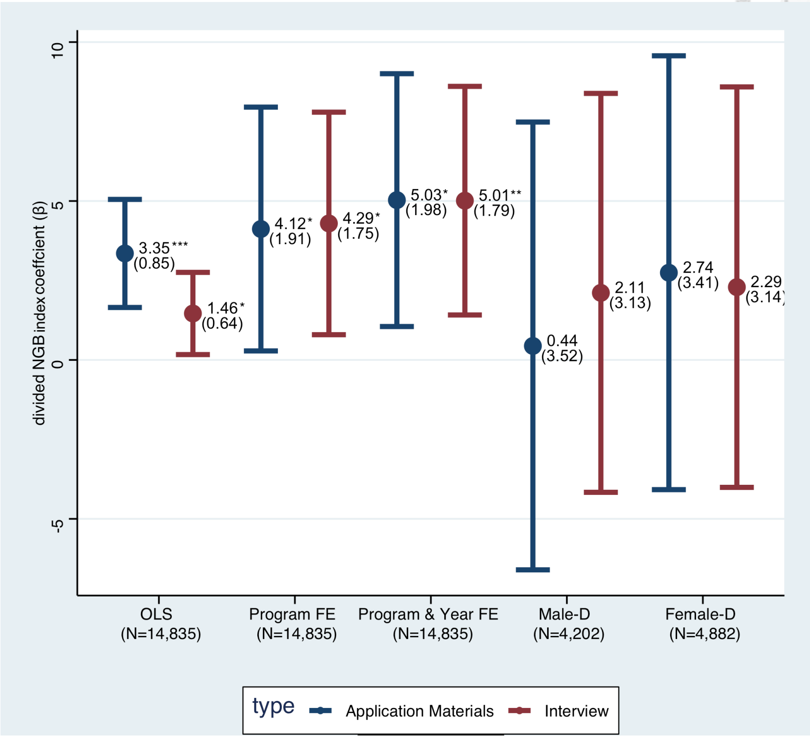

Figure A.7: Female evaluation advantage under non gender-blind tests (divided)(P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

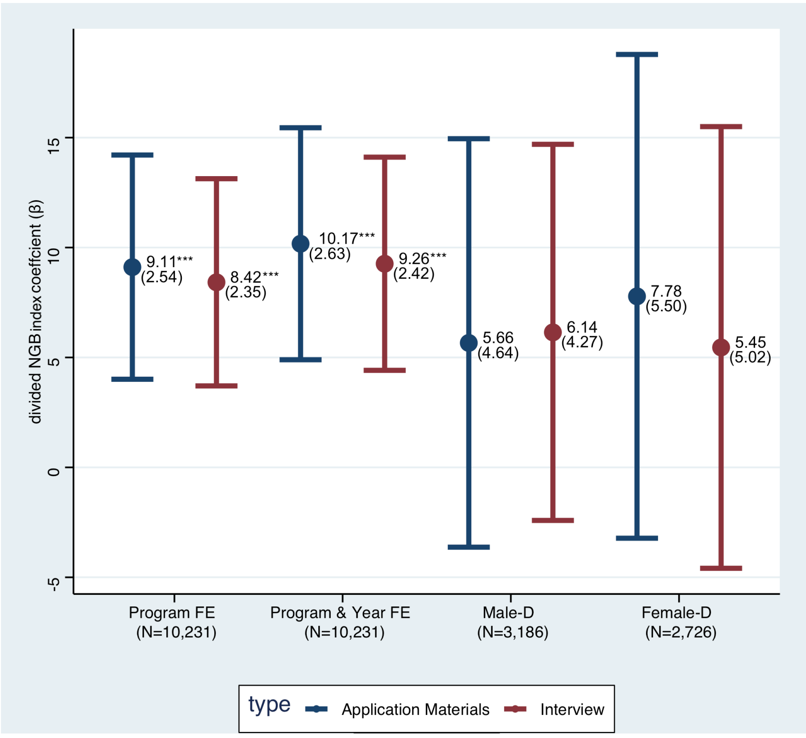

Figure A.8: Female evaluation advantage under non gender-blind tests (divided) w/o NMS and AH.(P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

Appendix VII: Analysis with different tiers

In Table 6, for tier 1, we have the schools whose 2021 QS rankings are within 300. Schools whose rankings are between 300 and 1000 are labeled with tier 2. The rest schools are further categorized in tier 3 to 5 by the following rules. Public schools outside QS rankings are in tier 3.[20] The remaining (private) schools are divided into half by their 2021 registration rates. Those with a rate higher than 82.5% are in tier 4 and those not are in tier 5.[21] In general, schools in tier 4 are usually considered the better private schools. The last column represents the ratio that the schools in each tier account for our data. One should notice that we do not include medical university and art university in this analysis; thus, the summation for the rest column will not be 100%. Note that National Chiao Tung University and National Yang-Ming University (both in the first tier) have merged into National Yang Ming Chiao Tung University in 2021.

Table A.1: School tiers by QS rank

Tier	QS rank	schools.	%
1	<300	National Taiwan University, National Tsinghua University, National Cheng Kung University, National Chiao Tung University, National Yang-Ming University	10.39
2	300-1000	National Taiwan Normal University, Taipei Medical University, National Sun Yat-Sen University, National Central University, Chang Gung University, National Chengchi University, National Chung Hsing University, Kaohsiung Medical University, National Chung Cheng University	13.96
3	>1000	the rest of the public schools	22.45
4	>1000	private schools with registration rate $\geq$ 82.5%	28.92
5	>1000	private schools with registration rate $<$ 82.5%	20.56

[20] In Taiwan, public schools are generally considered a better option than private schools because of lower tuition and abundant resource.

[21] For detail registration rate, see https://udb.moe.edu.tw/Index.

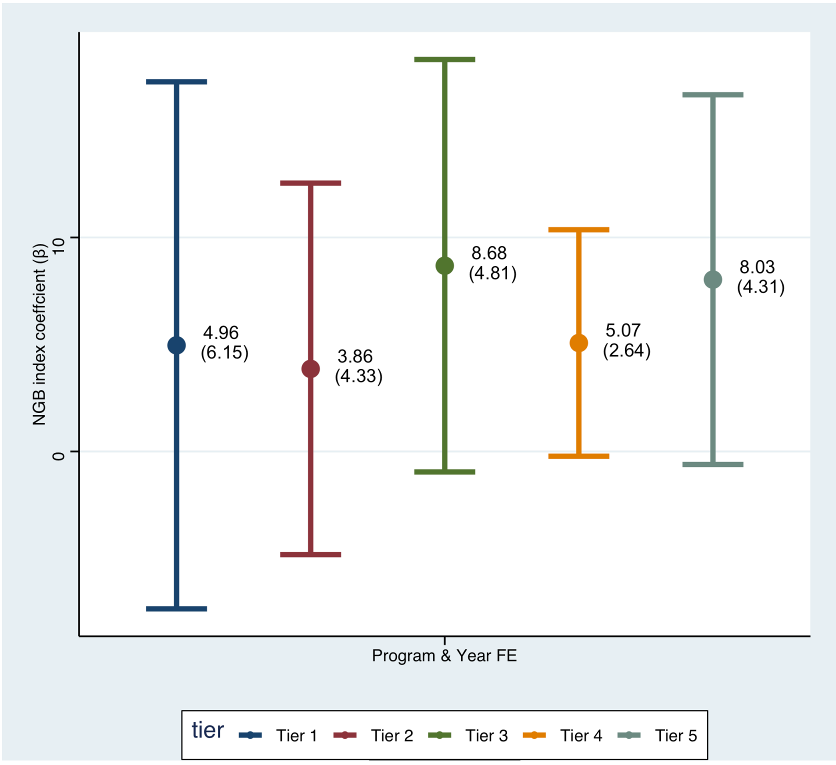

Figure A.9: Female evaluation advantage under non gender-blind tests in different tiers. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

Appendix VIII: Analysis weighted on program size

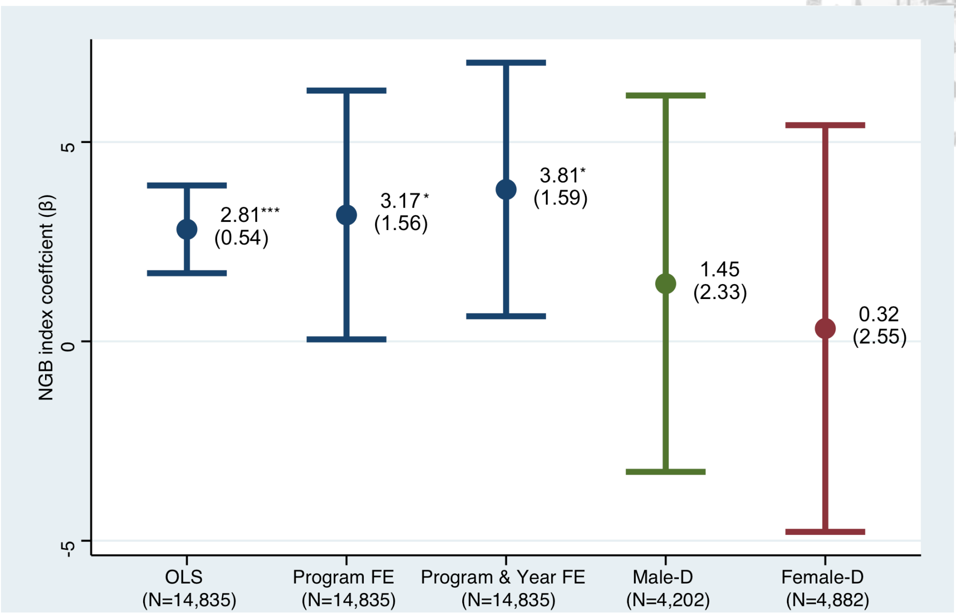

Figure A.10: Female evaluation advantage under non gender-blind tests weighted by program size. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

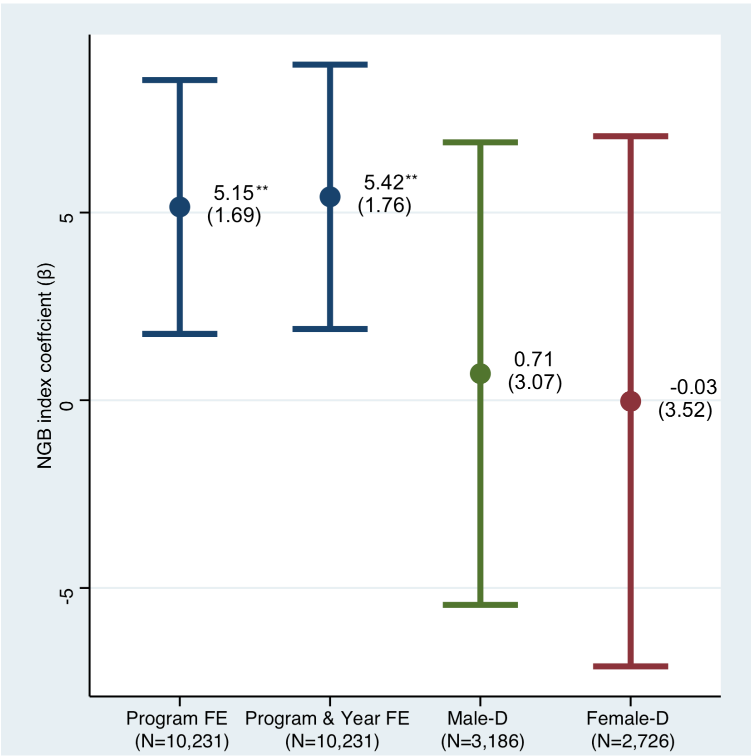

Figure A.11: Female evaluation advantage under non gender-blind tests weighted by program size w/o NMS and AH. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

Appendix IX: Analysis on other factors

Students are classified living in the urban area if the place they took the GSAT is either in Taipei, New Taipei, Taoyuan, Taichung, Tainan, or Kaohsiung which are the big cities in Taiwan.

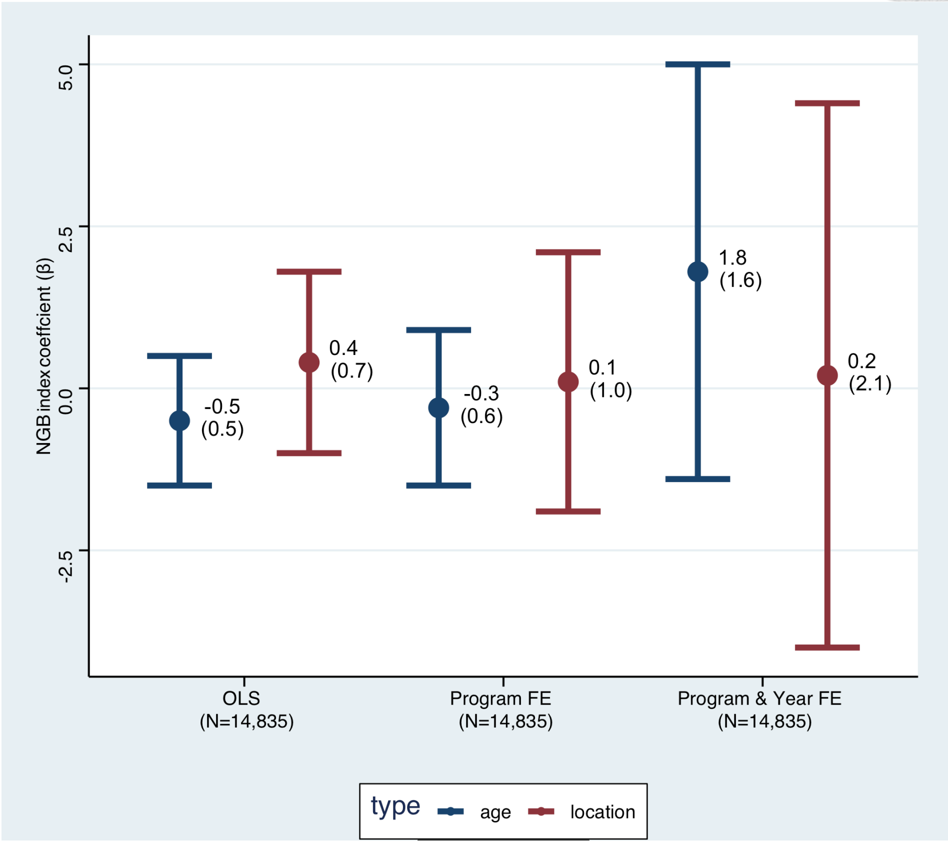

Figure A.12: Age and location distribution change under non gender-blind tests. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

The null effect may not be the single explanation for the above results since it might also be caused by the lack of variation for these two factors. Most high schools are located in the urban area in Taiwan; hence, over 70% of the pupils are classified as living in metropolitan in the data. Similarly, as we mentioned above, most applicants are at the same actual age, and the standard deviation for our name-predicted ages is only 3.8. To take a deeper look at this issue, we could compare the above results with the non gender-blind effect documented in specific gender-dominated disciplines, where these disciplines likewise have less sample variation. Although the lack of variation, a much larger effect size (although not statistically significant) is still observed in the forth and five columns from Figure 4 whereas the coefficient sizes are almost zero for age and location. Therefore, we believe the lack of variation could be a concern, yet a more reasonable explanation should be that non blind tests make no difference in terms of applicants' urban-rural distribution nor socioeconomic status.

Appendix X: Fixed-Effect on Different Levels

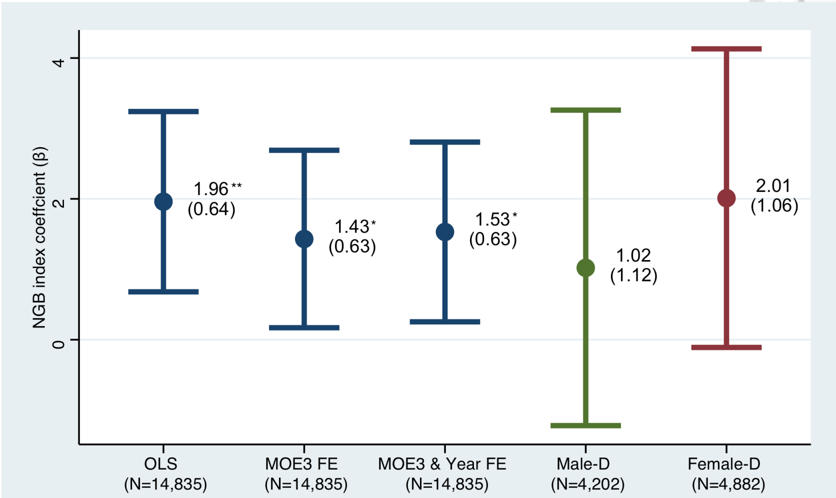

Figure A.13: Female evaluation advantage under non gender-blind tests fixed-effect on MOE 3. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)

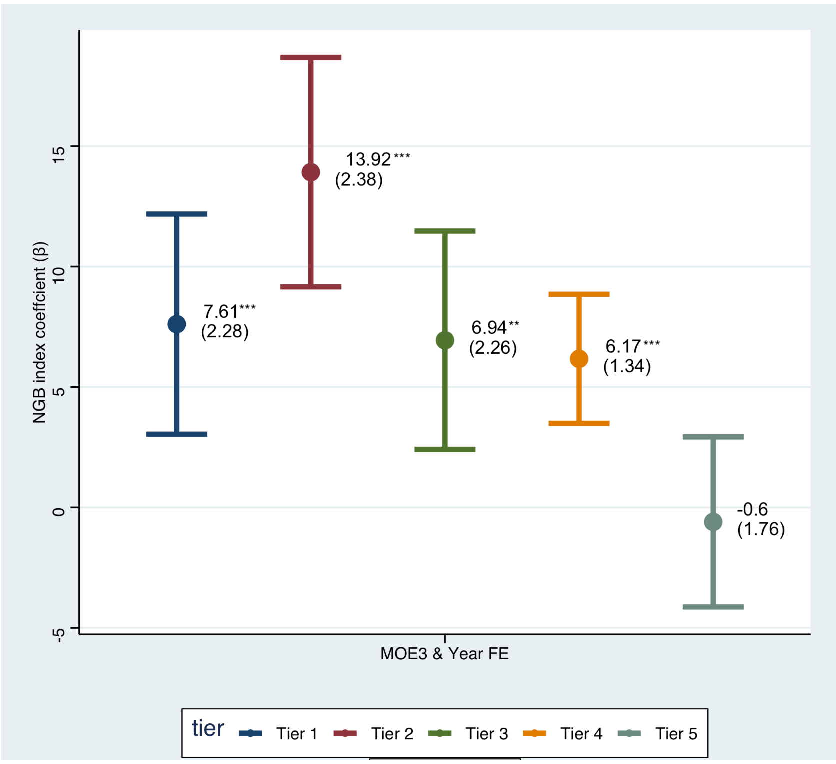

Figure A.14: Female evaluation advantage under non gender-blind tests fixed-effect on MOE 3 in different tiers. (P-value from Student's t-test and ***: p<0.001, **: p<0.01, *: p<0.05)