Public Sector Hiring and School Major Choices: Evidence from India

Shreya Chandra[*]

February 25, 2022

Abstract

High wage premiums for STEM fields, as well as substantial gender gaps in the choice of college major have been widely documented in the Economics literature. I document that the proportion of STEM graduates in India has nearly doubled from 2010 to 2020 - at least a decade after India's IT boom, which has been credited for driving similar changes in the late 1990s to early 2000s. In this paper, I explore the contribution of a positive labor demand shock for STEM graduates in a niche but highly prestigious public sector employer in India to this increase, exploiting a hiring policy change in 2011. I find that the hiring policy did indeed contribute to the increase, and that these increases are largely driven by two demographic groups - lower caste women, and upper caste urban residents.

[*] I would like to thank Jeremy Magruder for invaluable guidance through the course of this project. I would also like to thank Utkarsh Kumar and Shreya Sarkar for many helpful discussions.

1 Introduction

High wage premiums for STEM fields are widely documented in the Economics literature - Patnaik et al. (2020) estimate a 25-30% average earnings premium for STEM graduates in the US between 2016 and 2018.[1] Similarly in India, Jain et al. (2018) estimate a 22% STEM wage premium for men in 2012. India has also seen large increases in the proportion of STEM graduates at both the high school and university level, a change largely attributed to India's IT boom during the late 1990s and early 2000s - the proportion of STEM high school graduates in India grew from 15% in 1990 to 32% in 2010.[2,3] However, the gender gap in STEM high school graduates in India also increased from 2 to 11 percentage points during this period.[4] To the extent that major choices map to occupations and earnings, understanding what and who is driving these changes is crucial to understanding the determinants of income inequality. For example, Altonji et al. (2012) estimate that accounting for educational sorting by gender explains 54% of the gender wage gap in the US.

In this paper, I first document a further 20 percentage point increase in the proportion of STEM graduates among high school students in India by 2020 relative to 2010 using an event study design. This increase is largely driven by urban and upper caste individuals, with the gap between urban upper castes and rural, lower castes having increased from 4 percentage points before 2011 to roughly 9 percentage points after 2011. These trends are striking, and lend support to recent public agitation that policies that favor STEM graduates at hiring implicitly favor upper caste, urban, and "English speaking" citizens. For instance, the introduction of an aptitude test based screening process by one of India's most prestigious public sector employers in 2011 - the Union Public Service Commission (UPSC) - created

[1] These estimates are consistent with older estimates from the literature, as stated in Altonji et al. (2012).

[2] Own calculations from a backward looking age panel constructed using the IHDS 2011-2012 data.

[3] India is a unique setting where students select a "stream" or "major" as early as high school. Major choices in high school are typically followed through to college.

[4] Own calculations from a backward looking age panel constructed using the IHDS 2011-2012 data. In general, the gender gap in STEM graduates is well documented in India (Sahoo and Klasen, 2021), as it is in several other countries.

tremendous backlash among aspirants about favoring STEM. While this is a very niche part of the Indian labor market, public sector jobs are meant to be less discriminatory, and thus appeal to lower caste, rural citizens with promises of upward social mobility. This makes it important to understand who is driving the increase in STEM graduates, and what the labor market consequences of these choices are.

This motivates the second part of my paper, which estimates the contribution of the UPSC's 2011 policy in driving the increase in STEM graduates since 2010. I use variation by year and age at exposure to the policy - cohorts that were affected by the policy versus those just older - to estimate this effect. I estimate a large (almost double) increase in STEM uptake for lower caste women in 2011-12 relative to the control group in 2005-06 (the cohort of upper caste women that would not have been exposed to the policy). Previous literature on women's responses to changing labor market conditions in India finds conflicting evidence - on one hand, while STEM is becoming increasingly popular among female graduates in India, female labor force participation rates in India have fallen from 30.28% in 1990 to 20.52% in 2019.[5] This is puzzling for an economy whose growth has relied heavily on the high-skilled, services sector. Klasen and Pieters (2015) find that rising incomes for men, coupled with gender norms, result in a stronger income effect, particularly for urban, upper caste women, resulting in lower labor force participation rates for these women. On the other hand, lower caste women may be less constrained by gender norms and thus respond better to labor market conditions.[6] For instance, Munshi and Rosenzweig (2006) show that among cohorts that were exposed to India's IT boom in the 1990s, lower caste girls were more likely to attend English medium schools than lower caste boys, thus improving their future labor market outcomes.[7]

More broadly, it is not clear ex-ante whether the UPSC policy would affect individuals'

[5] World Bank Report

[6] Lower caste women tend to be poorer and have to work to earn their own wages, which may override gender norms about women's work, decision making and control over finances (Field et al., 2010).

[7] Munshi and Rosenzweig (2006) use an indicator for attending an English medium school as a proxy for better future labor market outcomes. Indeed, Patnaik et al. (2020) estimate a 22% English language premium for men in India, suggesting that this might be a meaningful proxy.

high school major choice, since candidates who choose to apply to the UPSC are themselves a highly self-selected group. Second, whether the UPSC policy really did favor STEM graduates, and how prevalent the public agitation towards this policy was are themselves empirical questions. I find that the UPSC policy increases the likelihood of choosing STEM among urban upper castes, and lower caste women, hinting at two very different potential mechanisms that drive this result. On one hand, if there was a response, the policy could have raised student's subjective expectations about the returns to choosing Science or altered their preferences towards different majors (including gender norms). This could explain why I find a larger response among lower caste women, who may value the non-pecuniary benefits of public sector more than men. This mechanism would be consistent with previous work on major choice and gender gap (Turner and Bowen, 1999; Zafar, 2013). On the other hand, the response may be driven by groups that are better able to respond to the policy along other dimensions such as income, access to better schooling infrastructure, etc.[8] This could explain why I find a large response among urban upper castes.

Finally, I also provide some suggestive evidence that the UPSC policy did not seem to favor STEM graduates. Figure A1 shows that the proportion of STEM graduates among UPSC admits has risen drastically over the past 20 years, which could easily translate into the public perception that the UPSC policy is favoring STEM graduates. Using a dataset on all UPSC admits from 1965 to 2020, I exploit variation by year and age at exposure to the policy to estimate whether among the pool of UPSC admits, individuals who were exposed to the policy were more likely to be a STEM graduate. I find that younger candidates (who were exposed to the policy) are in fact less likely (in some cases, significantly less likely) to be STEM graduates than their counterparts before the UPSC policy. However, due to both

[8] Higher secondary schools more expensive to build because they have hire requirements than lower grade schools in terms of teacher skills, equipment for labs (especially for Science), etc. Moreover, choosing Science as a major also involves cost considerations other than tuition and transportation costs. According to Jain et al. (2018), Science students are more likely to hire private tutors, the cost of which can amount up to 47% of total private education expenditure. It is also common for Science students in India to move away from their native city to cities that have "established themselves as national coaching hubs" for science students (Jain et al., 2018).

data limitations and the large extent of self-selection among UPSC admits, I am careful not to interpret this result as causal.

My paper contributes to two main strands of the literature. First, and the most direct contribution is to the literature on determinants of college major choice. Previous work has focused on quantifying the role of expected earnings, subjective expectations, preferences, and ability as determinants of major choice in developed country settings (Patnaik et al., 2020). However, there is little work in this space that exploits exogenous variation in the demand for particular majors. Studies that have exploited exogenous variation (using eligibility cut-offs along with regression discontinuity designs) have focused on estimating the returns to different majors conditional on choices, instead of estimating effects on choices themselves (Bertrand et al., 2010; Hastings et al., 2013; Kirkeboen et al., 2016; Andrews et al., 2017; Eckardt, 2020). My paper uses an exogenous demand shock for a particular major and directly estimates it's effects on major choice, in a developing country context. That India is one of the largest producers of STEM graduates in the world (cite OECD report) makes this setting particularly interesting. So far, there is little descriptive evidence on major choice, labor market consequences and income inequality in developing countries. Jain et al. (2018) examine the labor market effects of high school majors in India, and document a 21% wage premium for urban men, where the estimated returns are as high as 37% for the top 1% earners. Moreover, they find that these returns are mostly concentrated among the upper caste, consistent with my result that upper caste individuals, who might be better able to respond to the demand shock, are driving the increase in STEM graduates since 2010.

I also add to the literature on dynamic human capital investment decisions. For example, Khanna and Morales (2017) find that labor demand shocks in the US IT sector encouraged Indian college graduates to invest in IT skills. They leverage exogenous variation in the number of H-1B visas allotted to Indians to show that India experienced an overall "brain gain" even though the probability of migrating to the US was high, and led to the growth of the Indian IT sector. Along similar lines, Hahm and Park (2021) show how middle school choice in New York City public schools affects students High School choices. My paper is similar in that I estimate how individuals make decisions in high school based on a labor demand shock that will be realized at least 3 years into the future. In fact, my context explore the effects of a labor demand shock in the public sector, in combination with a seemingly strong public perception about the returns to STEM majors in a highly prestigious, albeit non-STEM occupation.

Finally, my paper contributes to the literature linking public sector hiring policies and aggregate labor market outcomes. This is particularly relevant in the Indian context, where public sector jobs command a 64% wage premium, conditional on region, occupation fixed effects and demographic controls (Finan et al., 2017). My paper is closest to Mangal (2021), which estimates the labor market consequences of a public sector hiring freeze in one state in India. Mangal (2021) finds that the hiring freeze is associated with a 30% lower labor force participation by the affected cohorts, who were induced to continue studying for the entrance examinations instead of switching occupations. This response came at a cost of lower earnings in the future, indicating a strong preference for public sector jobs among college graduates in India. My paper extends this work by focusing on a much larger, and more prestigious public employer - the Central government, and extends the analysis beyond one state in India.

The paper proceeds as follows: Section 2 discusses the institutional background and details of the UPSC policy. Section 3 describes data sources used. Section 4 discusses the estimation strategy, results are reported in Section 5 and Section 6 concludes.

2 Background

2.1 Higher Education in India

The education system in India requires students to choose a ”stream” in the higher-secondary, or post-secondary stage (i.e., when they enter grade 11). Students typically choose between Science, Business or Commerce and Humanities. Each major offers a (mostly) fixed set of courses that are designed to prepare students for college. Importantly, these choices are largely irreversible, and typically map one-to-one with the major they chose in college. Thus, unlike many other countries where students decide their majors before entering or during college, students in India make their major choices relatively early.

Students are admitted into high school majors based on school specific eligibility requirements, usually based on a cut-off score in the Secondary School Examinations (Jain et al., 2018). Cut-off scores for Science are typically higher than Business, followed by Humanities. In the Indian context, Science is also perceived to be a more desirable major choice, and so switching schools in order to meet the cut-off score for Science is not uncommon.

2.2 Public Sector Hiring Policies

The main policy variation I use in this paper to understand what contributed to the increase in Science majors is a positive demand shock for STEM graduates in a niche, but highly prestigious public sector organization. I focus on a change in the recruitment structure of the Civil Services examinations in India in 2011, administered by the Union Public Service Commission (UPSC), the body which recruits India's top-tier of bureaucrats. These officers go on to become Administrative, Foreign, Police and Revenue Service officers, positions that command a lot of prestige in society. Moreover, these jobs serve as a vehicle of upward mobility for candidates from lower socio-economic backgrounds as the examination based hiring process is meant to be less discriminatory than private sector jobs. The recruitment process is highly competitive - about a million individuals apply every year, spending 3-4 attempts (years) on average to clear the exam, while the acceptance rate is $< 0.01\%$. The process entails multiple stages - there are two examinations (the Prelims and the Mains), and one final interview round. Only about 5% of all applicants who appear for the Prelims exam make it to the next stage.[9]

[9] https://byjus.com/free-ias-prep/upsc-exam-success-rate-statistics-to-crack-the-exam-easily/

In 2011, the UPSC announced a change to the structure of the Prelims examination by replacing the existing subjective reasoning based test with a more objective aptitude test called the Civil Services Aptitude Test (CSAT).[10] The CSAT was designed as a logical reasoning test, with questions similar to Critical Reasoning tests in the GMAT, as opposed to the earlier format that asked more subjective reasoning based questions on topics such as Current Affairs, History, and Geography. The policy change was met with a lot of public dissatisfaction - aspirants viewed the change as favoring upper caste, urban and English speaking test takers, thus making the recruitment process unequal and discriminatory. Several protests (and an upcoming federal election) urged the government to later increase the maximum age limit for UPSC aspirants in 2014. This series of events indicates a strong preference for candidates to continue applying to these positions, instead of substituting away from them, and suggests that the policy may have had a response on major choice for individuals who were entering high school or college at the time of policy announcement. I exploit this variation in my estimation strategy to estimate the effect of this policy on high school stream choice.

This setting motivates two questions. First, given the public attention that this policy received, coupled with the prestige associated with public sector jobs, did the introduction of CSAT incentivize students to choose STEM majors in high school (or continue STEM in college)? Second, did the policy change disproportionately favor STEM graduates (.i.e., was it easier for STEM graduates to clear the UPSC examinations)? To answer the first question, I use a difference-in-difference and event study strategies to estimate the causal effect of introducing CSAT on high school stream choices in India using two separate sources of nationally representative secondary data sources. I also implement a difference-in-difference strategy to answer the second question using data on all Administrative Service officers *admitted* by the UPSC. While applicant level data would have been ideal to study the second question, I am currently limited by data on individuals who eventually cleared the

[10] Announcement made in 2010 to be in effect from 2011 Prelims.

UPSC examinations. However, my difference-in-difference analysis provides some suggestive evidence that even among the pool of admits, individuals were induced to switch to STEM in college after the introduction of CSAT. I am unable to distinguish between the effect of the CSAT exam itself from the effect of public perception.

3 Data

I use three datasets for my analysis, details of which are described below.

3.1 India Human Development Surveys

The India Human Development Surveys (IHDS) are a multi-topic, nationally representative household level survey collected by the National Council of Applied Economic Research. I use two rounds of the surveys, one conducted before the UPSC policy change in 2005-06 (IHDS I), and one just after the policy in 2011-12 (IHDS II). The IHDS is one of the few secondary datasets in India that asks information about high school or college major, as well as other questions about school characteristics, (school expenses and grants, test scores, etc.), along with a set of demographic and socio-economic characteristics. [11]

I use the IHDS data primarily for the difference-in-differences analysis to estimate the effect of CSAT on high school major choices. I restrict the sample to individuals who are between 16-19 years of age, ever attended school and have completed at least grade 10. This allows me to observe the set of individuals who ever made a decision about their high school field, both before and after the policy. This restriction yields 3,561 individuals in 2005-06, and 5,412 individuals in 2011-12.

I report descriptive statistics on the IHDS sample in Table 1. Individuals across both cross-sections are similar on most observables, with the exception of income, distance to school and the proportion attending public schools - all of which are higher in 2011-12 than

[11] The IHDS I (2005-06) codes all major choices into the three main categories: Science, Commerce or Humanities. The IHDS II (2011-12) data has more detailed value labels for major choice - for example, engineering is coded separately from science for individuals who are currently in college. For the purpose of my analysis, I recode the finer major choices into one of the three main stream choices.

2005-06. [12]

3.2 People of India (CMIE)

The People of India data (hereafter, CMIE data) is an individual level panel dataset collected by the Center for Monitoring the Indian Economy (CMIE). This dataset includes about 650,000 individuals surveyed over 3 waves each year from January 2014 to December 2020. The CMIE data is among the other few secondary data sources in India that collects information on field of study/field of work. Similar to IHDS II, I recode the finer major categories into the three categories of interest: Science, Commerce and Humanities. This data also contains detailed information on occupation (industry, income, time spent working, etc.), along with a set of demographic and socio-economic variables.

I use the last wave of the 2020 round, conducted between September and December of 2020, to construct a backward looking (unbalanced) panel of cohorts by the age at which individuals turned high school going age. That is, I code individuals who turned 16 years of age by the April of a particular year, with the value of that year. For instance, individuals who turned 16 in April 2020 are coded as Year = 2020, those who turned 17 by April 2019 as Year = 2019, and so on, as far back as year 2000. This allows me to estimate an event study specification to test for a change in the likelihood of choosing Science after the UPSC policy. Table A1 reports descriptive statistics for the CMIE data.

3.3 Supremo IAS Data

Finally, I use descriptive rolls of 6,451 Indian Administrative Service (IAS) officers in India from 1965 to 2020. These data were scraped from a publicly available executive record sheet of all IAS officers in India. For each officer, I observe some key demographic characteristics, such as age at entry to the Civil Services, educational qualification, field of study, caste, gender, marital status, and promotion history. I use this dataset to document changes in

[12] Income in IHDS surveys includes net income from agriculture and self-employment, which some households have reported to be negative. In all my current specifications using income as a control, I leave the variable as is.

the composition of IAS officers and empirically test whether the introduction of CSAT can be associated with the increase in Science graduates among the pool of IAS admits.Table 2 reports summary statistics for the Supremo data. The increase in the proportion of STEM graduates among the IAS admit is quite stark - STEM graduates in the IAS nearly triple after 2011.

4 Estimation Strategy

The main objective in this paper is to (1) document the increase in STEM graduates in India, (2) estimate the effect of introducing CSAT on high school major choice, and (3) provide suggestive evidence about whether the CSAT exams were associated with the increase in STEM graduates in the UPSC.

4.1 CSAT and High School Major Choice

First, I use an event-study specification to test whether the proportion of STEM graduates in India increased after the CSAT policy was announced. As mentioned before, I construct a backward looking panel of age-cohorts using the last available round of the CMIE, where the event year is the year in which an individual turned high-school going age. I run the following specification:

$$\text{Science}_{idt} = \alpha_0 + \sum_t \alpha_{1t}\mathbb{I}\{\text{Year} \leq 2009\}_{idt} + \sum_t \alpha_{2t}\mathbb{I}\{\text{Year} \geq 2011\}_{idt} + \mu_d + \epsilon_{it} \qquad (1)$$

where Science_{idt} is an indicator for choosing Science, or a STEM college major (engineering, medicine, etc.), $\mathbb{I}\{\text{Year} \leq 200X\}_{idt}$ is an indicator for whether individual i in district d at time t turned 16 years of age by April of that year, and μ_d denotes district fixed effects. Coefficients $\{\alpha_{1t}\}t$ test for parallel trends, and coefficients $\{\alpha_{2t}\}t$ estimate the increase in STEM uptake relative to 2010, one year before the CSAT policy was announced.

Further, to test for differential trends in STEM uptake by relevant social groups - gender, caste and region, I estimate the following double-difference event study specification:

$$\begin{aligned}\text{Science}_{idt} &= \alpha_0 + \sum_t \alpha_{1tG}\mathbb{I}\{\text{Year} \leq 2009\}_{idt} \times \mathbb{I}\{\text{Group}\}_{id} + \sum_t \alpha_{2tG}\mathbb{I}\{\text{Year} \geq 2011\}_{idt} \times \mathbb{I}\{\text{Group}\}_{id} \\ &+ \sum_t \alpha_{3t}\mathbb{I}\{\text{Year} \leq 2009\}_{idt} + \sum_t \alpha_{4t}\mathbb{I}\{\text{Year} \geq 2011\}_{idt} + \mathbb{I}\{\text{Group}\}_{id} + \mu_d + \epsilon_{it}\end{aligned} \tag{2}$$

where $\mathbb{I}\{\text{Group}\}_{id}$ is an indicator for whether individual i in district d belongs to one of the three groups: {Lower Caste, Female, Urban}. As a robustness check, I implement the same event study specification on IHDS-II (2011-12) - however, given the timing of this dataset, I am limited to only one post policy year.

I then use a difference-in-differences design to estimate the effect of the CSAT policy announcement on high school major choices using the IHDS data. Specifically, I exploit variation by age and time to identify this effect - in 2011, students entering high school (16-17 year olds) had the opportunity to strategically chose science in response to the CSAT policy announcement, while this choice had become irreversible for the just older cohorts (18-19 year olds), relative to the yeays before 2011. Thus, the first difference is by age cohort, where the treated group is 16-17 year olds (versus 18-19 year olds), and the second difference is time, where Post denotes observations in the IHDS II (versus IHDS I). I estimate this effect using the following specification:

$$\begin{aligned}\text{Science}_{idt} = \alpha_0 + \alpha_{\mathbf{1}} \text{ Post } 2011_t \times \text{ Treated}_{idt} + \alpha_2 \text{ Post } 2011_t \\ + \alpha_3 \text{ Treated}_{idt} + \mathbf{X_{idt}}\delta + \mu_d + \varepsilon_{idt}\end{aligned} \tag{3}$$

where all variables are as defined earlier, $\mathbf{X_{idt}}$ denotes a set of demographic and school controls, and α_1 is the coefficient of interest. As Munshi and Rosenzweig (2006) document, lower caste households may respond differently to education policies. I also use a triple difference strategy to test for differential effects by caste and gender using the following specification:

$$
\begin{aligned}
\text{Science}_{idt} = \alpha_0 + \alpha_{\mathbf{1}} \text{ Post } 2011_t \times \text{ Treated}_{idt} \times \mathbb{I}\{\text{Lower}\}_{id} \\
+ \alpha_2 \text{ Treated}_{idt} \times \mathbb{I}\{\text{Lower}\}_{id} + \alpha_3 \text{ Post } 2011_t \times \mathbb{I}\{\text{Lower}\}_{id} \\
+ \alpha_4 \text{ Post } 2011_t \times \text{ Treated}_{idt} + \alpha_5 \text{ Post } 2011_t + \alpha_6 \text{ Treated}_{idt} \\
+ \mathbf{X_{idt}}\delta + \mu_d + \varepsilon_{idt} \qquad (4)
\end{aligned}
$$

where $\mathbb{I}\{\text{Lower}\}_{id}$ is an indicator for individuals who belong to one of the three lower caste categories in India - Other Backward Caste (OBC), Scheduled Caste (SC), and Scheduled Tribes (ST). The remaining variables are defined as earlier. I also estimate Equation 3 and Equation 4 using the backward looking panel from the CMIE data, pooling all years after 2011 as Post 2011_t. Additionally, for both double and triple difference specifications, I test for heterogeneous effects by medium of instruction, gender and region (urban / rural) by running a fully saturated interaction of Pre/Post, Treatment, Caste and an indicator for medium of instruction, gender and region. I also run all specifications using an indicator for Science or Commerce, since students who chose Commerce also study mathematics, a core STEM subject.

4.2 Major Choice among IAS Admits

Finally, I present some descriptive evidence that the CSAT policy did indeed increase the number of science graduates among UPSC admits. In an ideal setting, I would have used applicant level data and tested (a) whether the proportion of STEM graduates among applicants increased post CSAT, and (b) whether CSAT really did disproportionately favor science graduates. [13]However, I am able to exploit the variation in exposure to UPSC policy by age and year among UPSC admits to provide some suggestive evidence for the impact for this policy. Age cohorts are considered to be treated if they were admitted to the UPSC after 2014, *and* were entering college in 2011 (the year the policy was announced).[14] This creates a staggered treatment of sorts, by age cohort and year: for instance, 21 year olds in 2014, who were entering college in 2011, are coded as ”Treated” in all years post 2014; 22 year olds are considered treated in 2015 and onwards, 23 year olds in 2016 and onwards, and so on. I limit my analysis to years before 2018, since Supremo only has one data point in 2019, and education data is missing for candidates in who entered in 2020, thus I only compare 21-25 year old cohorts before and after policy exposure. Given data limitations, I also restrict the sample for this regression to all years in which at least 80% observations have non-missing education data. I estimate a double difference regression, by treatment status and age - treatment status captures variation by time (pre-post policy exposure), and age captures variation across cohorts. I run the following specification:

$$\begin{aligned} \text{Science}_{iat} = \beta_0 &+ \sum_{a=21}^{24} \beta_{1a} \mathbb{I}\{\text{Treated}_{at}\} \times \mathbb{I}\{\text{Age} = \text{a}\} \\ &+ \sum_{a=21}^{24} \beta_{2a} \mathbb{I}\{\text{Age} = \text{a}\} + \beta_3 \mathbb{I}\{\text{Treated}_{at}\} + \mu_s + \varepsilon_{iat} \end{aligned} \tag{5}$$

where $\mathbb{I}\{\text{Treated}_{at}\}$ is an indicator for individuals who were exposed to the policy (ie entering college) in 2011, $\mathbb{I}\{\text{Age} = \text{a}\}$ is an age indicator and μ_s are state fixed effects. The base category is 25 year olds. Due to limited sample size, I also estimate Equation 5 by pooling across older age cohorts - I run two additional versions of this specification with different base categories: (1) pooling 21+ year olds ($a \in \{21, 21+\}$), and (2) pooling 22+

[13] A regression discontinuity design around the UPSC exam score cut-offs would enable me to test this.

[14] I define the post treatment year to be 2014 because college in India is typically 3-4 years long; so the first age cohort that was exposed to the policy could join the UPSC as early as 2014.

year olds ($a \in \{22, 22+\}$). [15] $\{\beta_{1a}\}$ are the coefficients of interest, which estimate the differential effect for treated cohorts after the policy.

4.3 Threats to Identification

4.3.1 Right to Education 2009

The passage of India's Right to Education (RTE) Act in 2009 poses a potential confounder to my proposed estimation strategy. Shah and Steinberg (2019) show that the RTE was associated with a 5 percentage point increase in school enrollment. However, in so far as the scheme was limited to enrollment up to grade 8, there is no obvious reason that this would affect how students chose major choices in grade 11. One channel could be that the compliers of the RTE may be differentially likely to chose a particular field, thus confounding the difference-in-difference results presented below. Another concern is that the RTE, by making 25% quotas for students from lower socio-economic background compulsory in private schools, may have set these students on a different trend than they were. However, given that students who chose STEM fields in high school are a highly selected group, it is unlikely that this might be the case. In future iterations of this paper, I aim to explicitly control for features of the RTE to isolate any confounding effects.

5 Results

5.1 Impact of CSAT on High School Stream Choice

Panel (a) in Figure 1 plots coefficients for each year from Equation 1 using the CMIE data, with robust standard errors clustered at the district level and 95% confidence intervals. The base year is 2010, one year before the announcement of CSAT. The coefficients before 2009 confirm the absence of any pre-trends in choosing Science as a major in high school. The proportion of STEM graduates has almost doubled in 2020 relative to 2010, off a base

[15] I choose to drop 21 year olds in the second version because there are four times as much fewer data points for that age group in the treatment period.

of 22% in 2010. Panels (b)-(d) break down these effects by Caste, Region, and Gender. I report estimates of the total change in STEM graduates for each group - i.e., for the base category, I report coefficients α_{4t}, and for the interaction term, I report the sum of coefficients $\alpha_{4t} + \alpha_{2Gt}$ from Equation 2. While there is no visible difference for men and women, the increase is significantly larger among urban and upper castes relative to their rural and lower caste counterparts. These trends are not surprising if we expect urban, upper castes to be best able to respond to the CSAT announcement. Table A2 reports results from the same event study specification, pooling across pre and post years in CMIE data - columns (1) reports the treatment effect over time (coefficient on Post) of 0.088 percentage points, which amounts to a 55% increase across years.

Next, I look at difference in difference estimates using the IHDS data. Since I only have one pre-policy round in the IHDS, I check for pre-trends by estimating the baseline difference in proportion of STEM graduates by treatment status (columns (1)-(3) in Table A3), as well as baseline difference-in-difference by caste and treatment status (columns (4)-(6)). Controlling for demographic variables eliminates any statistically distinguishable baseline difference in the proportion of STEM graduates (column (2) of Table A3). There seem to be no differential difference by caste at baseline. I further confirm this by estimating the event study specifications used on the CMIE data - Equation 1 and Equation 2 - using the backward looking panel of age cohorts using IHDS-II. The event study results from this regression are shown in panel(a) and (b) of Figure A3 - I am able to reject the presence of pre-trends in both the simple and double difference estimates.

Double Difference Estimates

Table 3 reports the double difference estimates from Equation 3 - columns (1)-(3) use an indicator for Science as the outcome variable, and columns (4)-(6) use an indicator for Science *or* Commerce as the outcome variable. Demographic and School level controls are added sequentially. The coefficient on Post × Treated is large and positive across all specifications

- treated cohorts are 0.03 percentage points (11.5%) more likely to choose STEM in high school than control cohorts before the policy. There doesn't seem to be any differential effect of CSAT on whether students went to an English medium school, or by gender and region: Table 4 reports heterogeneity results for the double difference specification by these three groups - English Medium (panel A), Female (panel B), and Rural (panel C). None of the triple interaction terms are significant, and I cannot reject F-tests of equality for Post × Treated × Group = Post × Treated across all specifications. However, the signs and magnitudes of these coefficients are robust to both sets of controls and are consistent with the perception that English schooled and urban students are more likely to respond to the CSAT policy by opting for STEM in high school. The sign of the Female interaction coefficient is ex-ante ambiguous.

Triple Difference Estimates

Next, I estimate the triple difference specification in Equation 4 to test for differential effects by caste. Columns (1)-(3) and (5)-(7) of Table 5 reports these results - I find no differential effect on the likelihood of opting for STEM by caste, and cannot reject that the coefficients by caste group are equal. Controlling for state level intensity of caste reservations does not change these estimates, suggesting that any potential effect is not being averaged out by differences in affirmative action policies.[16] These results are reported in columns (4) and (8) of Table 5. I also test for heterogeneity by interacting an indicator for Above Median and Below Median intensity states, reported in Table A4, but there is no significant difference by intensity of reservations. Finally, Table 6 reports heterogeneity results for the three groups - English Medium (panel A), Female (panel B), and Rural (panel C). I find no differential effects for lower castes by medium of instruction or region; however, I find large and statistically significant effects for lower caste women. The estimated effect sizes are almost as large as the control group mean (upper caste men in control cohorts in 2005).

[16] I follow the strategy adopted by Khanna (2020). Reservation intensity is defined as the proportion of Quota% for SCs, STs and OBCs combined, to the Population % in each state.

5.2 Impact of CSAT on Major Choice among IAS Admits

Finally, I present some suggestive evidence for the impact of CSAT on selection into the IAS, reported in Table 7. Columns (1)-(3) pool across all ages above 21, columns (4)-(6) pool across all ages above 22 (excluding 21 year olds), and columns (7)-(9) fully saturate on age dummies. As expected, the coefficient on Treated is large, positive, and significant in all specifications, reflecting the upward trend in STEM graduates among UPSC admits observed in the data (Figure A1). Conditional on being treated, 21 year olds are less likely to be STEM graduates (although the coefficients are not significant), while 22 and 23 year olds are more likely to be STEM graduates than their cohort counterparts before the CSAT policy. The magnitudes of these coefficients are robust to adding state fixed effects and demographic controls, although they become more imprecise.

6 Conclusion

I use a difference-in-differences design using two household level datasets to estimate the effect of the UPSC's CSAT policy on student's high school major choice in India. Based on anecdotal accounts, the introduction of CSAT created a lot of public uproar about favoritism towards STEM graduates, and implicitly, favoritism towards upper caste, urban test takers. I am also able to present some suggestive evidence that CSAT was not associated with an increase in STEM graduates among the pool of IAS officers, suggesting that public perceptions about favoritism towards STEM graduates may have been false. While only a very highly selected of individuals attempts the UPSC exam, the supposedly strong public opinion, along with promises of a stable, yet prestigious public sector job may still have encouraged students to choose STEM as a high school major in order to keep UPSC jobs in their future set of job options.

My estimates show a large increase in the proportion of STEM graduates since 2010, an increase that is largely driven by upper castes and urban students. While my results from the CMIE estimates indicate no differences by gender, estimates using the IHDS data indicate a large effect for the cohort lower caste women that was entering high school just after the policy announcement. Together, these results present two different potential mechanisms at work: first, that upper caste and urban citizens might be better able to respond to public perceptions, either through higher household incomes, access to better schooling infrastructure, or better access to information. Second, for lower caste women, the mechanism at play may relate more directly to individual preferences and the role of less restrictive gender norms than upper caste counterparts (Munshi and Rosenzweig, 2006). I plan to explore these mechanisms further in future versions of this paper. At the same time, I also acknowledge that these estimates are surprisingly large, given that the UPSC represents a very small segment of India's labor market. - I plan to investigate these effect sizes in future iterations.

References

J. G. Altonji, E. Blom, and C. Meghir. Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annu. Rev. Econ.*, 4(1):185–223, 2012.

R. J. Andrews, S. A. Imberman, and M. F. Lovenheim. Risky business? the effect of majoring in business on earnings and educational attainment. Technical report, National Bureau of Economic Research, 2017.

M. Bertrand, R. Hanna, and S. Mullainathan. Affirmative action in education: Evidence from engineering college admissions in india. *Journal of Public Economics*, 94(1-2):16–29, 2010.

D. Eckardt. *Training, occupations, and the specificity of human capital.* PhD thesis, London School of Economics and Political Science, 2020.

E. Field, S. Jayachandran, and R. Pande. Do traditional institutions constrain female entrepreneurship? a field experiment on business training in india. *American Economic Review*, 100(2):125–29, 2010.

F. Finan, B. A. Olken, and R. Pande. The personnel economics of the developing state. *Handbook of economic field experiments*, 2:467–514, 2017.

D. W. Hahm and M. Park. A dynamic framework of school choice: Effects of middle schools on high school choice. *Available at SSRN 3996418*, 2021.

J. S. Hastings, C. A. Neilson, and S. D. Zimmerman. Are some degrees worth more than others? evidence from college admission cutoffs in chile. Technical report, National Bureau of Economic Research, 2013.

T. Jain, A. Mukhopadhyay, N. Prakash, and R. Rakesh. Labor market effects of high school science majors in a high stem economy. *Available at SSRN 3286167*, 2018.

G. Khanna. Does affirmative action incentivize schooling? evidence from india. *Review of Economics and Statistics*, 102(2):219–233, 2020.

G. Khanna and N. Morales. The it boom and other unintended consequences of chasing the american dream. *Center for Global Development Working Paper*, 2017.

L. J. Kirkeboen, E. Leuven, and M. Mogstad. Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, 131(3):1057–1111, 2016.

S. Klasen and J. Pieters. What explains the stagnation of female labor force participation in urban india? *The World Bank Economic Review*, 29(3):449–478, 2015.

K. Mangal. Chasing government jobs: How aggregate labor supply responds to public sector hiring policy in india. *Working Paper*, 2021.

K. Munshi and M. Rosenzweig. Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy. *American Economic Review*, 96(4): 1225–1252, 2006.

A. Patnaik, M. J. Wiswall, and B. Zafar. College majors. Working Paper 27645, National Bureau of Economic Research, August 2020. URL `http://www.nber.org/papers/w27645`.

S. Sahoo and S. Klasen. Gender segregation in education: Evidence from higher secondary stream choice in india. *Demography*, 58(3):987–1010, 2021.

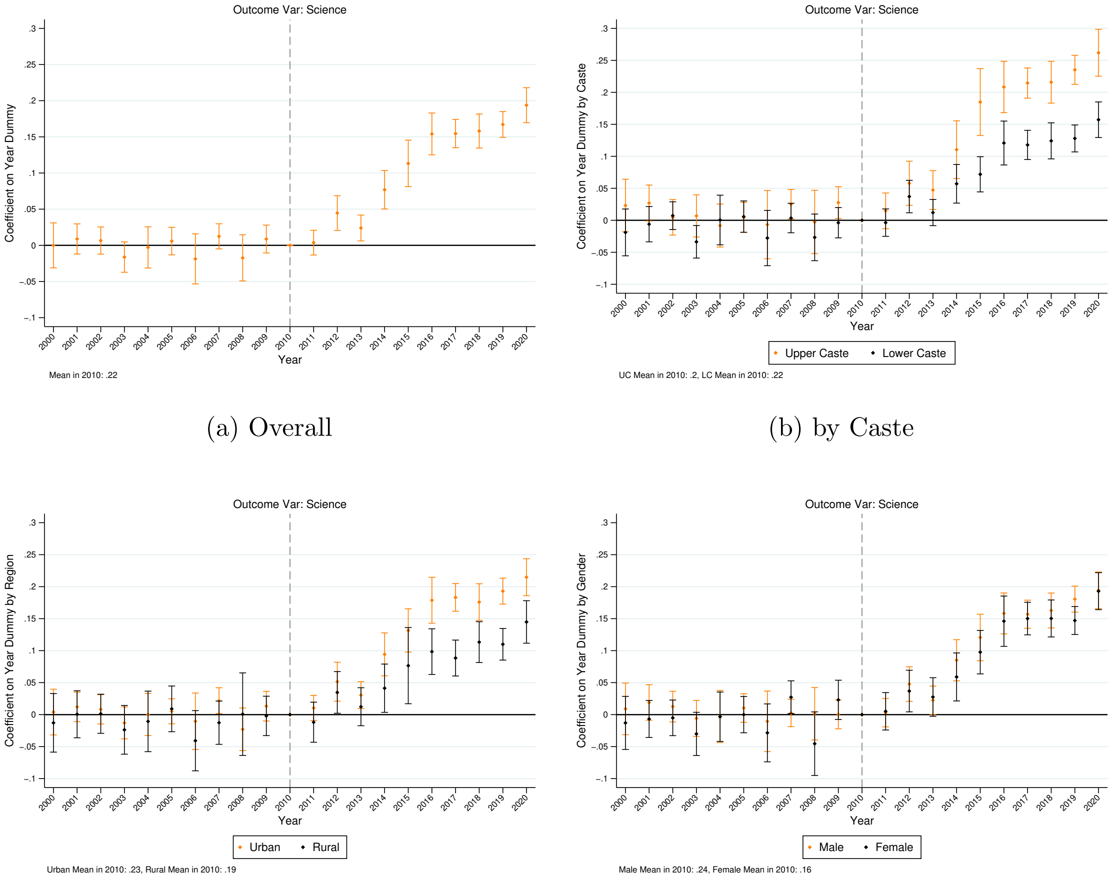
M. Shah and B. Steinberg. The right to education act: Trends in enrollment, test scores, and school quality. In *AEA Papers and Proceedings*, volume 109, pages 232–38, 2019.

S. E. Turner and W. G. Bowen. Choice of major: The changing (unchanging) gender gap. *ILR Review*, 52(2):289–313, 1999.

B. Zafar. College major choice and the gender gap. *Journal of Human Resources*, 48(3): 545–595, 2013.

Tables and Figures

Figure 1: Proportion of STEM Graduates since 2010 (CMIE)



(a) Overall

(b) by Caste

(c) by Region

(d) by Gender

Notes: These graphs report event study coefficients estimated from Equation 1 and Equation 2 using the CMIE data. Panel (a) plots coefficients for each year from Equation 1, with robust standard errors clustered at the district level and 95% confidence intervals. The base year is 2010, one year before the announcement of CSAT. Panels (b)-(d) break down these effects by Caste, Region, and Gender and plot estimates of the total change in STEM graduates for each group from Equation 2 - i.e., α_{4t} for the base category, and $\alpha_{4t} + \alpha_{2Gt}$ for the interaction term.

Table 1: Summary Statistics - IHDS

	IHDS I (2005-06) Mean/SD (1)	IHDS II (2011-12) Mean/SD (2)
Science	0.273	0.282
	(0.446)	(0.450)
Science or Commerce	0.416	0.456
	(0.493)	(0.498)
Female	0.472	0.499
	(0.499)	(0.500)
Age	17.813	17.902
	(0.972)	(1.014)
Urban	0.479	0.574
	(0.500)	(0.495)
Lower Castes (ST, SC)	0.185	0.227
	(0.389)	(0.419)
Lower Castes (SC, ST, OBC)	0.512	0.574
	(0.500)	(0.494)
HH Sise	5.808	5.672
	(2.638)	(2.522)
Household Assets (Rs)	16.842	18.858
	(5.393)	(5.450)
Household Income (Rs.)	83315.975	174072.909
	(86605.828)	(311163.404)
Highest level of education (Females)	7.271	7.058
	(5.017)	(5.134)
Distance to School (km)	7.721	8.652
	(8.469)	(11.529)
Government School	0.172	0.266
	(0.378)	(0.442)
Observations	3561	5412

Notes: Standard deviations in parentheses. This table reports summary statistics for the restricted sample I use for my analysis - individuals who are between 16-19 years of age, ever attended school and have completed at least grade 10.

Table 2: Summary Statistics - Supremo IAS Data

	Overall	Pre CSAT	Post CSAT
	1965 -2020	1965 - 2010	2011-2020
	Mean/SD	Mean/SD	Mean/SD
	(1)	(2)	(3)
Number of Admits	115.196	106.043	157.300
	(40.173)	(29.512)	(55.767)
Age at Last Exam	25.166	24.745	26.472
	(2.636)	(2.458)	(2.740)
Male	0.822	0.848	0.743
	(0.382)	(0.359)	(0.437)
Married (Women only)	0.003	0.004	0.003
	(0.058)	(0.061)	(0.050)
Has Science Degree	0.292	0.216	0.626
	(0.455)	(0.411)	(0.484)
Has Tech Degree	0.160	0.107	0.395
	(0.367)	(0.309)	(0.489)
Has Medical/Architechture Degree	0.047	0.033	0.110
	(0.212)	(0.178)	(0.314)
Has Bachelors Degree	0.993	0.991	0.998
	(0.086)	(0.093)	(0.043)
Has Masters Degree	0.592	0.636	0.397
	(0.492)	(0.481)	(0.490)
Elite College	0.117	0.072	0.310
	(0.321)	(0.259)	(0.463)
Observations	6451	4878	1573

Notes: Standard deviations in parentheses. Pre CSAT changes refers to years before 2011, before *any* changes to the UPSC exam were made. CSAT (Civil Services Aptitude Test) refers to the inclusion of logical reasoning style questions to the UPSC prelims examination in 2011. Marital Status is only available for women. Number of admits is the average number of admits across all years in the sample, which runs from 1965-2020. Year 2019 has only one observation available. Age at last exam is the exact age of the candidate one year before the candidate was selected (i.e., at the time of application for the last attempt).

Table 3: Double Difference Estimates

	Y = Science			Y = Science or Commerce		
	(1)	(2)	(3)	(4)	(5)	(6)
Post × Treated	0.030	0.028	0.034*	0.018	0.019	0.027
	[0.020]	[0.020]	[0.020]	[0.023]	[0.020]	[0.020]
Treated	0.049**	-0.016	-0.018	0.055***	0.027	0.025
	[0.022]	[0.029]	[0.029]	[0.020]	[0.031]	[0.031]
Post	0.002	-0.017	-0.015	0.041**	0.017	0.021
	[0.014]	[0.013]	[0.013]	[0.018]	[0.017]	[0.017]
Constant	0.254***	0.866***	0.990***	0.392***	0.627**	0.774***
	[0.009]	[0.173]	[0.168]	[0.011]	[0.238]	[0.238]
Control Group Mean (2005-06)	.26	.26	.26	.39	.39	.39
R-Squared	.03	.07	.08	.04	.13	.14
Demographic Controls		X	X		X	X
School Controls			X			X
Observations	8973	8973	8973	8973	8973	8973

Notes: This table reports double difference estimates from Equation 3 using both IHDS-I and IHDS-II data. Standard errors in brackets and clustered at district level. Treated is an indicator for individuals who are 16-17 years of age, versus those who are just older - 18/19 years. Post is an indicator for individuals observed after 2011 (IHDS-II). All regressions include district fixed effects. Demographic controls include household size, income, assets, highest female education, gender and age. School controls include type of school (public/private) and distance to school. Control group refers to individuals in the control group (18-19 year olds) observed before the policy (2005).

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

Table 4: Double Difference Estimates - Heterogeneity

	Y = Science			Y = Science or Commerce		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: English Medium						
Post × Treated × English Medium	0.057	0.060	0.050	0.029	0.030	0.017
	[0.070]	[0.071]	[0.070]	[0.078]	[0.081]	[0.077]
Post × Treated	0.023	0.021	0.029	0.015	0.016	0.026
	[0.023]	[0.022]	[0.022]	[0.027]	[0.024]	[0.024]
Constant	0.243***	0.776***	0.888***	0.380***	0.538**	0.674***
	[0.010]	[0.174]	[0.169]	[0.011]	[0.240]	[0.240]
Control Mean	.24	.24	.24	.38	.38	.38
(p-value) Post × Treated × Group = Post × Treated	0.444	0.429	0.472	0.713	0.739	0.801
R-Squared	.06	.09	.1	.07	.14	.15
Panel B: Female						
Post × Treated × Female	-0.019	-0.021	-0.017	-0.023	-0.026	-0.020
	[0.052]	[0.052]	[0.051]	[0.052]	[0.054]	[0.054]
Post × Treated	0.039	0.040	0.044	0.027	0.034	0.039
	[0.029]	[0.029]	[0.030]	[0.031]	[0.031]	[0.031]
Constant	0.276***	0.861***	0.984***	0.433***	0.624*	0.771**
	[0.016]	[0.173]	[0.168]	[0.015]	[0.238]	[0.238]
Control Mean	.28	.28	.28	.43	.43	.43
(p-value) Post × Treated × Group = Post × Treated	0.972	0.854	0.914	0.917	0.786	0.863
R-Squared	.04	.07	.08	.05	.13	.14
Panel C: Rural						
Post × Treated × Rural	-0.066	-0.058	-0.055	-0.034	-0.023	-0.021
	[0.051]	[0.049]	[0.049]	[0.050]	[0.046]	[0.047]
Post × Treated	0.068	0.059	0.064	0.039	0.031	0.037
	[0.036]	[0.034]	[0.035]	[0.029]	[0.026]	[0.027]
Constant	0.309***	0.866***	0.992***	0.495***	0.624*	0.774**
	[0.017]	[0.172]	[0.167]	[0.021]	[0.239]	[0.239]
Control Mean	.31	.31	.31	.5	.5	.5
(p-value) Post × Treated × Group = Post × Treated	0.172	0.233	0.225	0.781	0.987	0.997
R-Squared	.04	.07	.08	.08	.13	.14
Demographic Controls		X	X		X	X
School Controls			X			X
Observations	8973	8973	8973	8973	8973	8973

Notes: This table reports heterogeneity in double difference estimates from Equation 3 using both IHDS-I and IHDS-II data. Standard errors in brackets and clustered at the district level. Treated is an indicator for individuals who are 16-17 years of age, versus those who are just older - 18/19 years. Post is an indicator for individuals observed after 2011 (IHDS-II). All regressions include district fixed effects, and demographic and school controls. Demographic controls include household size, income, assets, highest female education, gender and age. School controls include type of school (public/private) and distance to school. Control group refers to individuals in the control group (18-19 year olds) observed before the policy (2005), and in the base category (non-English medium, Males or Urban).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Triple Difference Estimates

	Y = Science				Y = Science or Commerce			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post × Treated × LC	0.014	0.015	0.019	0.010	0.007	0.005	0.009	0.005
	[0.037]	[0.038]	[0.038]	[0.038]	[0.047]	[0.047]	[0.047]	[0.048]
Treated × LC	0.011	0.004	0.003	0.013	0.009	-0.001	-0.002	0.004
	[0.032]	[0.033]	[0.032]	[0.032]	[0.037]	[0.038]	[0.037]	[0.037]
Post × LC	-0.033*	-0.035*	-0.036*	-0.029	0.002	-0.003	-0.003	0.003
	[0.018]	[0.019]	[0.019]	[0.020]	[0.029]	[0.028]	[0.028]	[0.028]
Post × Treated	0.021	0.018	0.023	0.027	0.013	0.016	0.022	0.026
	[0.024]	[0.026]	[0.027]	[0.028]	[0.033]	[0.033]	[0.033]	[0.035]
Treated	0.043*	-0.019	-0.019	-0.018	0.050**	0.027	0.026	0.032
	[0.023]	[0.027]	[0.027]	[0.027]	[0.024]	[0.032]	[0.032]	[0.032]
Post	0.020	-0.003	-0.000	-0.014	0.043	0.015	0.018	0.000
	[0.017]	[0.019]	[0.019]	[0.020]	[0.027]	[0.027]	[0.027]	[0.027]
Lower Caste (LC)	0.017	0.069***	0.067***	0.048***	-0.034	0.045*	0.044*	0.028
	[0.016]	[0.016]	[0.017]	[0.018]	[0.022]	[0.023]	[0.022]	[0.024]
Constant	0.245***	0.817***	0.942***	1.011***	0.410***	0.592**	0.740***	0.819***
	[0.013]	[0.169]	[0.164]	[0.166]	[0.015]	[0.239]	[0.239]	[0.238]
Control Mean	.25	.25	.25	.25	.42	.42	.42	.42
(p-value) Post × Treated × LC = Post × Treated	0.906	0.957	0.954	0.784	0.933	0.878	0.872	0.796
R-Squared	.03	.08	.08	.08	.04	.13	.14	.13
District FE	X	X	X		X	X	X	
Demographic Controls		X	X	X		X	X	X
School Controls			X	X			X	X
State Reservation Intensity				X				X
Observations	8973	8973	8973	8973	8973	8973	8973	8973

Notes: This table reports triple difference estimates from Equation 4 using both IHDS-I and IHDS-II data. Standard errors in brackets and clustered at the district level. Treated is an indicator for individuals who are 16-17 years of age, versus those who are just older - 18/19 years. Lower caste includes OBCs, SCs and STs. Post is an indicator for individuals observed after 2011 (IHDS-II). All regressions except columns (4) and (8) include district fixed effects. Demographic controls include household size, income, assets, highest female education, gender and age. School controls include type of school (public/private) and distance to school. Control group refers to upper caste individuals in the control group (18-19 year olds) observed before the policy (2005). Column (4) includes an additional control for intensity of reservations for lower castes at the state level.

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

Table 6: Triple Difference Estimates - Heterogeneity

	Y = Science			Y = Science or Commerce		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: English Medium						
Post × Treated × LC × English Medium	0.053	0.098	0.097	-0.032	0.031	0.030
	[0.164]	[0.158]	[0.162]	[0.176]	[0.176]	[0.180]
Post × Treated × LC	-0.004	-0.019	-0.017	0.044	0.014	0.015
	[0.064]	[0.065]	[0.067]	[0.072]	[0.072]	[0.075]
Constant	0.235***	0.735***	0.831***	0.393***	0.568	0.694**
	[0.022]	[0.274]	[0.264]	[0.023]	[0.355]	[0.347]
Control Mean	.24	.24	.24	.4	.4	.4
(p-value) Post × Treated × LC × Group = Post × Treated × LC	0.632	0.809	0.826	0.457	0.670	0.687
R-Squared	.08	.12	.12	.09	.16	.17
Panel B: Female						
Post × Treated × LC × Female	0.270**	0.216*	0.216*	0.226*	0.151	0.148
	[0.088]	[0.088]	[0.088]	[0.101]	[0.107]	[0.106]
Post × Treated × LC	-0.154*	-0.131	-0.131	-0.098	-0.074	-0.073
	[0.074]	[0.074]	[0.078]	[0.091]	[0.090]	[0.092]
Constant	0.294***	0.853**	0.961***	0.471***	0.685	0.820*
	[0.025]	[0.284]	[0.272]	[0.025]	[0.363]	[0.353]
Control Mean	.28	.28	.28	.47	.47	.47
(p-value) Post × Treated × LC × Group = Post × Treated × LC	0.028	0.095	0.105	0.135	0.407	0.449
R-Squared	.06	.1	.11	.08	.16	.16
Panel C: Rural						
Post × Treated × LC × Rural	-0.030	-0.070	-0.071	0.002	-0.049	-0.053
	[0.116]	[0.112]	[0.116]	[0.115]	[0.116]	[0.117]
Post × Treated × LC	0.007	0.027	0.027	0.007	0.030	0.031
	[0.064]	[0.066]	[0.066]	[0.080]	[0.082]	[0.081]
Constant	0.279***	0.832**	0.944***	0.483***	0.658	0.798*
	[0.023]	[0.280]	[0.267]	[0.030]	[0.362]	[0.352]
Control Mean	.29	.29	.29	.5	.5	.5
(p-value) Post × Treated × LC × Group = Post × Treated × LC	0.555	0.825	0.834	0.542	0.931	0.954
R-Squared	.06	.1	.11	.1	.16	.16
Demographic Controls		X	X		X	X
School Controls			X			X
Observations	8973	8973	8973	8973	8973	8973

Notes: This table reports heterogeneity in triple difference estimates from Equation 4 using both IHDS-I and IHDS-II data. Standard errors in brackets and clustered at the district level. Treated is an indicator for individuals who are 16-17 years of age, versus those who are just older - 18/19 years. Lower caste includes OBCs, SCs and STs. Post is an indicator for individuals observed after 2011 (IHDS-II). All regressions include district fixed effects, and demographic and school controls. Demographic controls include household size, income, assets, highest female education, gender and age. School controls include type of school (public/private) and distance to school.

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

Table 7: Composition of STEM Graduates among UPSC Admits

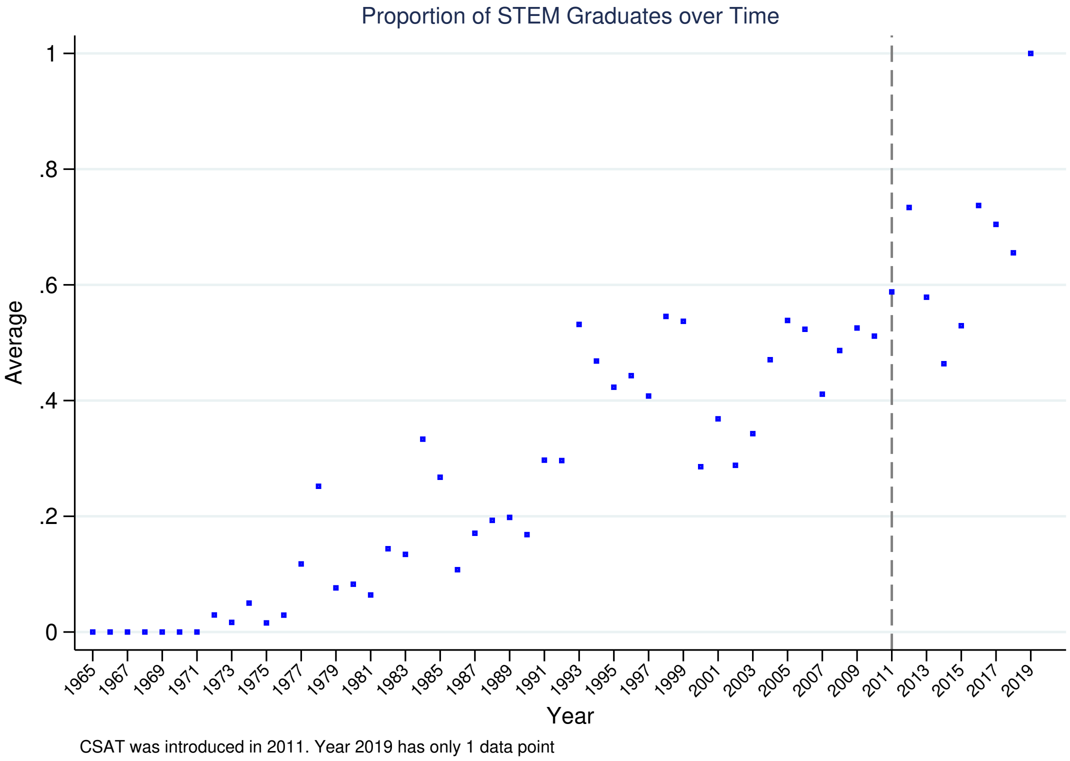
	Y = STEM Graduate								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treated × Age = 21	-0.252	-0.237	-0.242				-0.337*	-0.309	-0.310*
	[0.163]	[0.213]	[0.174]				[0.180]	[0.224]	[0.178]
Treated × Age = 22				-0.048	-0.045	-0.047	-0.122	-0.107	-0.105
				[0.093]	[0.099]	[0.099]	[0.119]	[0.137]	[0.136]
Treated × Age = 23							-0.069	-0.047	-0.034
							[0.107]	[0.109]	[0.105]
Treated × Age = 24							-0.187	-0.178*	-0.190**
							[0.122]	[0.103]	[0.093]
Age = 21	0.008	0.004	0.011				0.036**	0.031*	0.042**
	[0.015]	[0.011]	[0.010]				[0.017]	[0.018]	[0.017]
Age = 22				0.044***	0.039	0.041*	0.060***	0.055*	0.058**
				[0.014]	[0.024]	[0.021]	[0.015]	[0.032]	[0.028]
Age = 23							0.028*	0.025	0.031
							[0.015]	[0.024]	[0.021]
Age = 24							0.035**	0.035**	0.035**
							[0.017]	[0.015]	[0.015]
Treated	0.048	0.052	0.029	0.053	0.057	0.037	0.111	0.103	0.075
	[0.050]	[0.031]	[0.033]	[0.055]	[0.038]	[0.038]	[0.091]	[0.074]	[0.078]
Constant	0.289***	-0.002	-0.177***	-0.015***	-0.016	-0.197***	-0.036***	-0.034	-0.219***
	[0.005]	[0.014]	[0.020]	[0.005]	[0.025]	[0.031]	[0.011]	[0.035]	[0.038]
Control Group Mean	.29	.29	.29	.32	.32	.32	.3	.3	.3
R-Squared	.27	.28	.32	.26	.27	.31	.27	.28	.32
State Fixed Effects		X	X		X	X		X	X
Controls									
Observations	5903	5903	5903	5426	5426	5426	5903	5903	5903

Notes: This table reports results from Equation 5. Standard errors in brackets. All regressions include year fixed effects. Demographic controls include controls for caste (general category dummy), gender, whether the candidate has a masters degree and whether they went to an elite university. Treated is an indicator for individuals who were exposed to the policy (i.e. entering college) in 2011, and $\mathbb{I}\{\text{Age} = \text{a}\}$ is an age indicator. The control group is ages 22+ in columns (1) - (3), 23+ in columns (4) - (6), and 25 in columns (7) - (9).

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

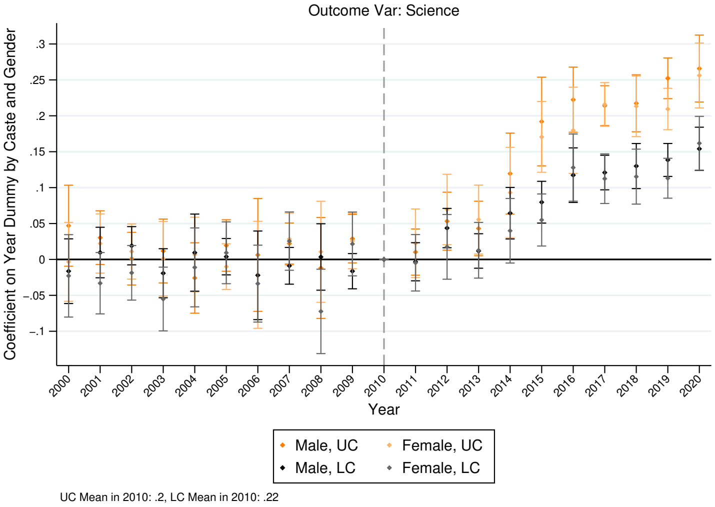
Appendix

Figure A1: STEM Graduates in UPSC Over Time

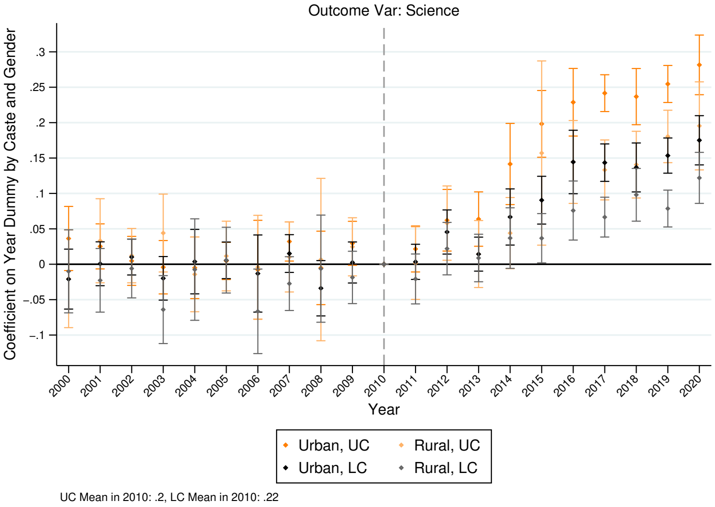


Notes: This graph reports the proportion of Science graduates by year among the pool of IAS officers using the Supremo IAS dataset.

Figure A2: Proportion of STEM Graduates by Caste since 2010 (CMIE)



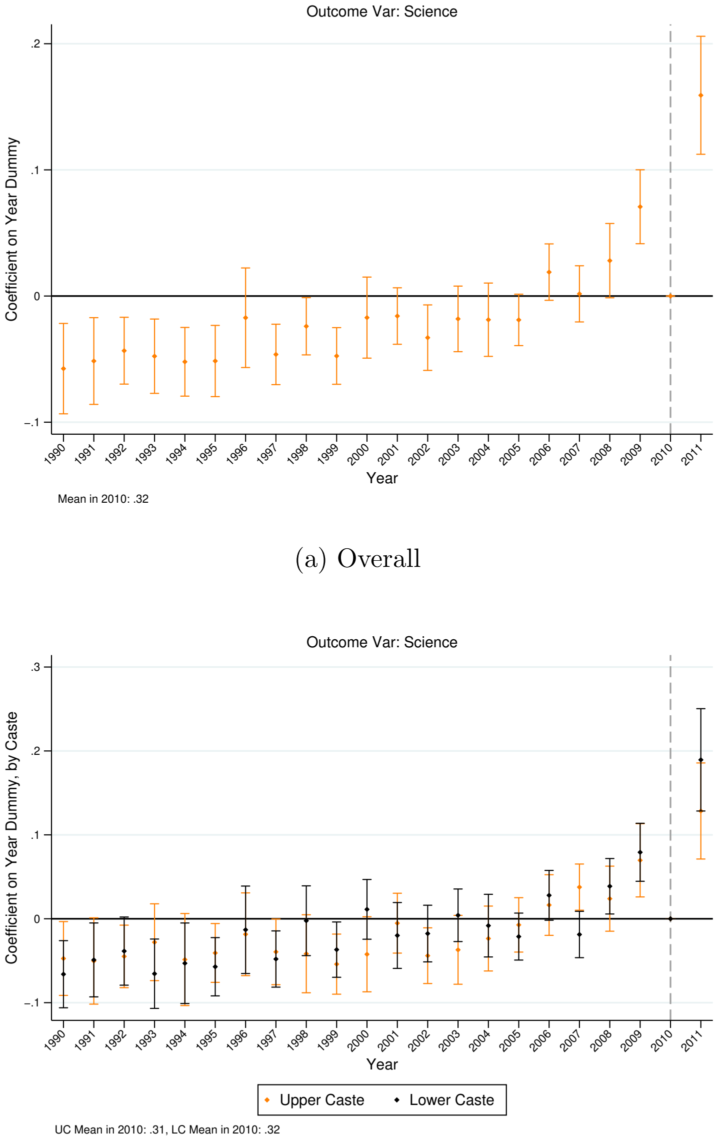
(a) Caste × Gender



(b) Caste × Region

Notes: These graphs report event study coefficients estimated from Equation 2 using the CMIE data, with fully saturated interactions for each Group. Panel (a) plots the total effect for each of the four groups in the interaction of Caste and Gender, and Panel (b) plots the total effect for the interation of Caste and Region.

Figure A3: Proportion of STEM Graduates by Caste since 2010 (IHDS)



(a) Overall

(b) by Caste

Notes: These graphs report event study coefficients estimated from Equation 1 and Equation 2 using the IHDS-II data. Panel (a) plots coefficients for each year from Equation 1, with robust standard errors clustered at the district level and 95% confidence intervals. The base year is 2010, one year before the announcement of CSAT. Panel (b)breaks down these effects by Caste and plots estimates of the total change in STEM graduates for lower and upper castes from Equation 2 - i.e., α_{4t} for the base category(upper castes) and $\alpha_{4t} + \alpha_{2Gt}$ for lower castes.

Table A1: Summary Statistics - CMIE

	Round 2020	
	Pre 2010 Mean/SD	Post 2010 Mean/SD
	(1)	(2)
Science	0.162	0.322
	(0.369)	(0.467)
Science or Commerce	0.283	0.466
	(0.450)	(0.499)
Female	0.458	0.371
	(0.498)	(0.483)
Age	32.187	20.238
	(3.061)	(2.565)
Rural	0.315	0.325
	(0.464)	(0.468)
Lower Castes (ST, SC)	0.246	0.262
	(0.431)	(0.440)
Lower Castes (ST, SC, OBC)	0.653	0.661
	(0.476)	(0.473)
HH Sise	4.497	4.481
	(1.610)	(1.321)
Household Income (Rs.)	20926.303	20901.890
	(16533.142)	(17202.554)
Highest level of education (Females)	9.843	10.036
	(2.297)	(3.140)
Observations	27735	64482

Notes: Standard deviations in parentheses. This table reports summary statistics for CMIE survey respondents who turned 16 from the year 2000 to 2020, and completed grade 10 at the time of the survey. These data are from the third wave of the 2020 CMIE round, conducted between September and December 2020. Column (1) reports results for the control group (i.e. those who turned 16 before 2010), and column (2) reports results for the treatment group.

Table A2: Difference-in-Difference Estimates (CMIE)

	Outcome Var: Science Major					
	Pre-Post	Double Difference			Triple Difference	
	Post	Post × Caste	Post × Gender	Post × Region	Post × Caste × Gender	Post × Caste × Region
	(1)	(2)	(3)	(4)	(5)	(6)
Post × LC × Female					0.005	
					[0.015]	
Post × LC × Rural						0.009
						[0.015]
Post × LC		-0.017**			-0.019*	-0.016*
		[0.008]			[0.010]	[0.010]
LC × Female					-0.002	
					[0.014]	
LC × Rural						-0.004
						[0.013]
Post × Female			0.003		0.000	
			[0.008]		[0.012]	
Post × Rural				-0.025***		-0.027**
				[0.009]		[0.013]
LC		-0.022***			-0.021**	-0.020**
		[0.007]			[0.009]	[0.008]
Post	0.088***	0.098***	0.087***	0.095***	0.098***	0.103***
	[0.006]	[0.007]	[0.007]	[0.007]	[0.009]	[0.008]
Female			-0.003		-0.004	
			[0.008]		[0.011]	
Rural				-0.037***		-0.034**
				[0.011]		[0.014]
Constant	0.181***	0.194***	0.182***	0.191***	0.196***	0.203***
	[0.005]	[0.005]	[0.006]	[0.006]	[0.008]	[0.007]
Control Mean	.16	.14	.16	.17	.13	.15
R-Squared	.23	.23	.23	.23	.23	.23
p-val Post × Group = Post		0	0	0		
p-val Post × LC × Group = Post × LC					0	0
Observations	88891	88891	88891	88891	88891	88891

Notes: This table reports double different estimates from Equation 3 using the CMIE 2020 data. Robust tandard errors in brackets and clustered at the district level. Post is an indicator for individuals who entered high school after 2011. Lower caste includes OBCs, SCs and STs. All regressions include district fixed effects.

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

Table A3: Pre Trends Check (IHDS)

	Difference by Age			Double Difference by Caste × Age		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Y = Science						
Treated × LC				0.016	0.009	0.008
				[0.032]	[0.033]	[0.033]
Treated	0.052**	0.019	0.021	0.044*	0.012	0.015
	[0.022]	[0.038]	[0.038]	[0.023]	[0.032]	[0.033]
Lower Caste (LC)				0.020	0.070***	0.070***
				[0.016]	[0.016]	[0.016]
Constant	0.255***	0.556*	0.595**	0.245***	0.513*	0.552**
	[0.007]	[0.283]	[0.280]	[0.011]	[0.276]	[0.275]
Control Mean	.26	.26	.26	.26	.26	.26
R-Squared	.04	.07	.08	.04	.08	.08
Observations	3561	3561	3561	3561	3561	3561
Panel B: Y = Science or Commerce						
Treated × LC				0.011	0.002	0.001
				[0.038]	[0.039]	[0.038]
Treated	0.056***	0.058	0.060	0.050**	0.056	0.059
	[0.020]	[0.038]	[0.038]	[0.025]	[0.038]	[0.038]
Lower Caste (LC)				-0.033	0.042*	0.043*
				[0.023]	[0.022]	[0.022]
Constant	0.397***	0.336	0.390	0.414***	0.310	0.364
	[0.007]	[0.318]	[0.317]	[0.012]	[0.313]	[0.313]
Control Mean	.39	.39	.39	.41	.41	.41
R-Squared	.04	.13	.13	.04	.13	.13
Demographic Controls		X	X		X	X
School Controls			X			X
Observations	3561	3561	3561	3561	3561	3561

Notes: This table reports results from regressions that test for baseline differences in the likelihood of choosing STEM (or difference-in-differences by Caste) using IHDS I (2005-06). Standard errors in brackets and clustered at the district level. Treated is an indicator for individuals who are 16-17 years of age, versus those who are just older - 18/19 years. Lower caste includes OBCs, SCs and STs. Panel A reports results for the outcome variable *Science*, which is an indicator for choosing science in high school. Panel B reports results for *Science or Commerce*, which is an indicator for choosing science or commerce in high school. All regressions include district fixed effects. Demographic controls include household size, income, assets, highest female education, gender and age. School controls include type of school (public/private) and distance to school.

Table A4: Triple Difference Estimates - Heterogeneity by Intensity in Caste Reservations

	Y = Science			Y = Science or Commerce		
	(1)	(2)	(3)	(4)	(5)	(6)
Post × Treated × LC × Above Median Intensity	-0.016	-0.017	-0.024	0.065	0.056	0.049
	[0.117]	[0.114]	[0.116]	[0.120]	[0.113]	[0.114]
Post × Treated × Above Median Intensity	0.132	0.112	0.115	0.092	0.075	0.078
	[0.084]	[0.081]	[0.080]	[0.082]	[0.078]	[0.078]
Treated × Above Median Intensity	-0.073	-0.071	-0.075	0.001	0.000	-0.004
	[0.062]	[0.060]	[0.061]	[0.058]	[0.051]	[0.052]
Post × Above Median Intensity	0.016	0.001	-0.002	0.015	0.004	0.001
	[0.053]	[0.050]	[0.049]	[0.056]	[0.052]	[0.052]
Treated × LC × Above Median Intensity	-0.054	-0.050	-0.049	-0.102	-0.097	-0.098
	[0.085]	[0.081]	[0.083]	[0.089]	[0.080]	[0.081]
Post × LC × Above Median Intensity	0.027	0.021	0.026	-0.015	-0.030	-0.025
	[0.053]	[0.054]	[0.054]	[0.070]	[0.066]	[0.066]
LC × Above Median Intensity	-0.094*	-0.098*	-0.098**	-0.077	-0.071	-0.072
	[0.052]	[0.049]	[0.048]	[0.054]	[0.048]	[0.047]
Post × Treated × LC	-0.004	-0.012	-0.010	-0.007	-0.022	-0.021
	[0.098]	[0.096]	[0.099]	[0.091]	[0.086]	[0.090]
Treated × LC	0.063	0.054	0.057	0.079	0.073	0.078
	[0.081]	[0.077]	[0.078]	[0.080]	[0.074]	[0.076]
Post × LC	-0.044	-0.047	-0.048	0.012	0.014	0.013
	[0.043]	[0.041]	[0.041]	[0.051]	[0.048]	[0.047]
Post × Treated	-0.051	-0.044	-0.040	-0.068	-0.057	-0.052
	[0.070]	[0.068]	[0.070]	[0.060]	[0.056]	[0.059]
Treated	0.083	0.010	0.012	0.064	0.025	0.025
	[0.061]	[0.062]	[0.063]	[0.050]	[0.053]	[0.054]
Post	0.030	0.000	0.003	0.046	0.004	0.007
	[0.045]	[0.041]	[0.041]	[0.046]	[0.041]	[0.040]
Lower Caste (LC)	0.046	0.107**	0.105**	-0.022	0.059	0.057
	[0.046]	[0.042]	[0.041]	[0.048]	[0.044]	[0.043]
Constant	0.337***	0.953***	1.047***	0.507***	0.795**	0.914***
	[0.040]	[0.273]	[0.264]	[0.039]	[0.339]	[0.334]
Control Mean	.33	.33	.33	.52	.52	.52
(p-value) Post × Treated × LC × Group = Post × Treated × LC	0.421	0.480	0.447	0.884	0.914	0.868
R-Squared	.1	.15	.16	.11	.2	.21
Observations	8973	8973	8973	8973	8973	8973

Notes: This table reports heterogeneity in triple difference estimates from Equation 4 using both IHDS-I and IHDS-II data. Standard errors in brackets and clustered at the district level. Treated is an indicator for individuals who are 16-17 years of age, versus those who are just older - 18/19 years. Lower caste includes OBCs, SCs and STs. Post is an indicator for individuals observed after 2011 (IHDS-II). All regressions include district fixed effects, and demographic and school controls. Demographic controls include household size, income, assets, highest female education, gender and age. School controls include type of school (public/private) and distance to school. Control group refers to upper caste individuals in the control group (18-19 year olds) observed before the policy (2005), and in the base category (below median reservation intensity, non-English medium, Males or Urban).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$