Interplay between Education and Identity: Inter Caste Marriages, Gendered Stream Choice and Caste Peer Effects

Komal Sahai

Thesis submitted to the Indian Statistical Institute in partial fulfilment of the requirements for the degree of Doctor of Philosophy

To my parents

Interplay between Education and Identity: Inter Caste Marriages, Gendered Stream Choice and Caste Peer Effects

Komal Sahai

January 2021

Thesis Supervisor: Dr. Tridip Ray

Thesis submitted to the Indian Statistical Institute in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Acknowledgements

I am indebted to my thesis supervisor, Dr. Tridip Ray, for his invaluable guidance and support. I have been extremely fortunate to have worked under his able supervision. He has been a source of great inspiration for me and I have learnt much from him. I always look up to him for his ability to critically analyse text and subtext and his openness to new ideas. He has been a guardian figure and I feel lucky to have developed a unique bond with him over these years.

I cannot thank enough Dr. Arka Roy Chaudhuri, who has been like an elder brother and a coach to me. He is the person I would turn to for any big and small query, and I could always rely on his expertise. I value his kindness and his honesty. I am grateful to him for having faith in me and for always looking out for me.

I would like to extend my gratitude to all the faculty members at the Economics and Planning Unit, Indian Statistical Institute, Delhi, for their teaching, support and cherished memories. I am grateful to Dr. Bharat Ramaswami and Dr. Abhiroop Mukhopadhyaay for their thoughtful suggestions and helpful comments. I thank Dr. Farzana Afridi, Dr. Prabal Roy Chowdhury and Dr. E. Somanathan for their encouragement and interest in my work. I am also immensly thankful for the staff at ISI: Mr. Anil Shukla, Ms. Deepmala, Babulal ji, Jagdish ji, (late) Pappu bhaiya, Rajendar bhaiya and Sunaina ji and all the mess staff. They are the silent cogs in the institute's machinery and they certainly made my time at ISI much more easy, delightful and memorable.

I am especially thankful to the Indian Statistical Institute for providing an ideal environment for research. I deeply cherish the time spent at the ISI hostel and the research lab, where work, discussion and chatter went hand in hand. The unique

climate at the institute brings together students who help each other during the ups and downs of an academic degree. I have found some lifelong friends in Gursharn, Dyotona di, Sujaya, Ahana, Nikita, Ankit, Swagatam, Ritabrata and Midhu. I would like to especially mention Prachi di and Dhritiman bhaiya for their love, care, innumerable helps and uncountable fun times.

I feel it is impossible to finish one's PhD thesis without the support of family. My parents Digambar Sahai and Sumati Sahai have been my unwavering wall of support and encouragement. I am what I am because of them. I am eternally blessed to always have the company of my brother, Shikhar, during my best and my worst times. His presence can uplift me from any abyss. My friends Sayantika and Priyamvada are my sisters from other mothers. I cannot imagine a day without them. I cherish my friendship with Charuhas and Suraj. I thank Divyani for believing in me when I myself did not. I am immensely grateful to have Sonal, Charu and Sakshi as my closest friends and confidantes.

I cannot skip to mention my dog and my best friend, Bruno from this list. His innocent eyes, adorable games and his mere presence around me have pulled me through the lowest of days.

Last, but not the least, I thank my husband and life partner, Manas. I cannot thank him enough for tirelessly taking care of me and everything around me every day. His encouragement and timely humour have helped me through the day even during the most stressful times. I am extremely lucky to have his love and friendship in my life.

Contents

Chapter 1

Introduction

This thesis lies at the intersection of two broad themes in economics: education and identity. Education is a primary tool for building human capital in economies. It has been recognised as the driver of economic prosperity via research and innovation (Barro 2001; Aghion et al. 2009; Hanushek and Woessmann 2010). At the same time, education is also touted as the driver of social progress and institutional change. A more educated society fares better on indices of equality (Gylfason and Zoega 2003), health (Ross and Wu 1995; Cutler and Lleras-Muney 2006) and political awareness and participation (Mayer 2011).

However, looking at education as an instrument of social change naturally brings in the equation the concept of social identity or simply identity. A social identity can be defined as a religious, racial, biological or geographical group with which an individual identifies herself. The three chapters in this thesis are dedicated to gaining an understanding about the interplay of education and identity in the context of India.

One of the primary dimensions of identity in the Indian society is caste (Desh-

pande 2011; Munshi 2016). Caste is a system of social stratification based on the ancient Hindu *varna* system which divided groups based on their occupations (Deshpande 2011). These *varnas* are determined by birth and are mutually exclusive and non-changeable. The Indian Human Development Survey, a nationally representative survey, shows that almost every person in the country, irrespective of her religion, identifies herself as belonging to a caste, even though it was originally a Hinduism concept. The pervasiveness and the continued relevance of caste as a component of identity in India makes it a compelling candidate to study.

The first broad aspect I examine about the relationship between education and identity is if and how education is associated with caste based practices. One of the central features of the institution of caste is caste endogamy (Bidner and Eswaran 2015). Endogamy serves the purpose of upholding caste boundaries and violations of this custom often invites punishment and social ostracism (Kaur 2010). The structure is extremely rigid and adherence is remarkably high even in modern India. Out-marriages, or inter caste marriages in this context, is a rare occurrence. In my first chapter, I examine if education can bring about a change in this centuries old practice of marrying within one's own caste, keeping in mind the "arranged marriage" aspect in the Indian society where marriages are mostly fixed by parents of the spouses.

Another major dimension of identity in India, and in general, is gender. The fundamental biological nature of this dimension makes it a basic identity trait in any human society. More importantly, a divergence in socioeconomic resources and outcomes along gender lines is also a reality in most societies (Giuliano 2020). Gender divide in educational outcomes is one such area of academic interest. While there is undoubtedly a gender gap in educational achievements, like literacy rate and en-

rollment rate (Census, 2011), a more curious aspect is when girls and boys studying alongside each other make vastly different educational choices. In particular, students in India have to choose specialized subjects of study after their matriculation in school and, interestingly, these choices show stark gendered patterns. This is the focus of the second chapter. It looks at the gender divide in subject choices of students after matriculation and explores the possible factors that can account for this gap.

The last chapter revisits caste as the dimension of identity but asks how caste identity might affect educational outcomes instead of the other way round. Access to education in India has historically been an elite privilege (Cheney et al. 2005). In the past, only the upper castes of *Brahmins*, the priests and teachers, and *Kshatriyas*, the warriors and protectors, enjoyed any form of education. The majority of the masses, consisting of castes lower in the hierarchy, remained uneducated (Deshpande 2011; Hnatkovska et al. 2012). This changed after Independence in 1947. Spreading education to the masses became a priority of the state of India (Cheney et al. 2005). On paper, this meant that an individual from any caste, religion or gender could get an education. To ensure that this was also true in practice, the Constitution of India implemented a system of affirmative action. It reserved 22.5% seats for persons belonging to the historically marginalised castes (called the Scheduled Castes or SC) and tribes (called the Scheduled Tribes or ST) of India in political constituencies, government jobs and higher educational institutions.

However, an inherent prejudice or apprehension against people from disadvantaged backgrounds is still a reality in many places (*Hindustan Times* (New Delhi, 19 September, 2012); *News 18* (Saharanpur, 14 April, 2018)). People believe that the presence of students from lower caste families or poor families in general can have

a negative impact on the quality of education of other students via, perhaps, lower class performance or increased indiscipline. The third chapter of my thesis explores this question in a peer effects framework. It examines the effects of peers belonging to scheduled castes and tribes on the academic performance of students.

Below I provide a brief overview of each of the chapters, outlining the research questions, empirical strategies and the results.

1.1 Whose education matters? An analysis of inter caste marriages in India

Caste endogamy is central to the institution of caste which has been shown to be discriminatory (Shah 1985; Thorat and Newman 2007) and detrimental to democracy (Jeffrey 2002; Munshi 2017), social mobility (Munshi and Rosenzweig 2006; Munshi 2017), trade (Anderson 2011) and environment (Gadgil and Rao 1994). Violations of the endogamy norm are often punished by social ostracism (Kaur 2010). It is also one of the most resilient caste based practices till date. The rate of inter caste marriages, even as recent as in 2011, was as low as 5.82% and there has been no upward time trend over the past four decades. Caste endogamy is the pillar of caste system and inter caste marriages can directly weaken the foundations of this system. This chapter studies the relationship of caste endogamy with education, which, according to Dr. B.R. Ambedkar, could free the marginalized sections of the society and set them on a path of upward mobility (Velaskar 2012; Moon and Narke 2014a,b). It takes into account the nature of the Indian marriage market where marriages arranged by parents and close relatives is largely the norm.

We use a nationally representative dataset, the second round of the Indian Human

Development Survey (IHDS-II), to quantify this relationship. But, at the outset, we recognize that we have to pay due attention to the fact that marriage markets in India work very differently as compared to the Western countries. A majority of marriages are arranged by the parents, and the spouses barely know each other before marriage. For example, 73% of marriages in our dataset were reported to have been arranged by parents and almost 70% of the women said that they met their husbands only on the day of their wedding.

The wide prevalence of the arranged marriage institution in the Indian marriage markets strongly suggests that any analysis of marriages in India must consider parental attributes along with individual ones. To justify this approach, we first explore whether education levels of the spouses themselves have any predictive power on the likelihood of inter caste marriages. We find that, contrary to the findings in the existing literature on out-marriages in the West, especially in the USA, the education levels of the individuals themselves do not have any association with the probability of inter caste marriages. The result is very robust to the inclusion of a whole range of controls and fixed effects, and to variations in the sample. We also attempt to disentangle our null results to see if they mask opposing effects of education via different mechanisms or if they mask heterogeneity across caste groups. We find no evidence for either case.

Having established the irrelevance of the spouses' own education, we next explore whether parental education is associated with the likelihood of an inter caste marriage. We add the education levels of the parents of both the spouses to our set of explanatory variables. Here we find that the level of education of the husband's mother has a positive and statistically significant association with the likelihood of an inter caste marriage. One standard deviation increase in the years of education of

the husband's mother is associated with a 10.16% increase in the probability of inter caste marriages over the sample mean. The result is very robust to variations in the sample, to the addition of a number of controls as well as fixed effects, to alternate model specification and to omitted variable bias. However, this part of the result is nuanced in the sense that among the parents on both sides, only the education of the husband's mother has a predictive power on the likelihood of inter caste marriage. We posit some potential channels based on theoretical arguments from the existing literature and provide some suggestive evidence for our proposed mechanism.

1.2 What can(not) explain the gap? Evidence and Decomposition of Gendered Stream Choice in India

Gender gap in earnings is well established in the economics literature, both in the context of developed (Blau and Kahn 2017; Boll and Lagemann 2018) as well as developing countries (Chi and Li 2014; Guimarães and Silva 2016). While a number of explanations have been extended to explain this gap, occupational segregation has emerged as a major explanatory factor (Hegewisch et al. 2010; Hegewisch and Hartmann 2014; Blau and Kahn 2017). In particular, male dominated occupations related to Science, Technology, Engineering and Mathematics (STEM) fields have substantial earnings premium while women are over represented in lower paying jobs like nursing and teaching (Webber 2016; Belfield et al. 2018; Dahl et al. 2020). Why are so few women employed in STEM related occupations despite a clear economic advantage in these fields? It could be either because fewer women graduate in STEM or because women graduate in STEM but drop out of STEM occupations, or a

combination of both. In most countries, however, specialization into STEM, or more broadly, into Science and non-Science streams happens even earlier, at the school level itself. In this chapter, we look at the very first stream choices made by students at the school level in the context of India using the results data from the Central Board of Secondary Education.

We first establish and quantify the gender gap in stream choice. We find a clear gender divide in our dataset along the same lines as observed in the literature. For example, on an average, boys are 19.13 percentage points more likely than girls to take-up Mathematics in class XI and 20.61 percentage points more likely to take-up the Physics-Chemistry-Mathematics (PCM) combination. Girls, on the other hand, are 11.18 percentage points more likely than boys to take-up Biology. In general, a higher proportion of girls takes up Biology, Economics, Political Science and History, while boys are more likely to take-up PCM and Computer Science.

Next we proceed to dissect this gap. Using regression and linear decomposition techniques, we decompose this gap to estimate how much of it can be "explained" by an expansive set of explanatory factors that have been proposed in the economics, sociology and psychology literature. In particular, we evaluate how much of the gender gap is accounted for by a difference in student ability, their cohort peers, their immediate seniors and their socioeconomic status. We conclusively show that a difference in student ability, the earliest and most common explanation offered, accounts for less than 10% of the gender gap we observe in various subjects. A difference in cohort peer attributes also does not explain any statistically significant portion of the gap. Instead, we propose a novel way to use the immediate seniors of students to elicit measures of role model and "chilly" climate. We propose that a student's seniors may serve as potential role models and that students may also form

an idea about how friendly or hostile the environment will be in a particular subject by looking at the gender composition of the subject class of their seniors. We find that these role model and chilly climate aspects of a student's immediate seniors in school are the largest explanators of the existing gender gap in stream choice. If girls had the share of own gender students in the senior Mathematics and PCM classes like boys, and had the share of own gender seniors choosing Biology like boys, the gender gap in these subjects would have closed by 24%, 16% and 18%, respectively.

1.3 Caste peer effects on student performance: Evidence from Indian schools

Existence of peer effects in education is a common wisdom that is becoming an increasingly rigorously proven fact in the economics of education literature (Hoxby 2000; Hanushek et al. 2003; Angrist and Lang 2004; Billings et al. 2014; Antecol et al. 2016). Peer attributes can be peer ability and performance, peer background or peer identity like gender or race. In this chapter, we use three cohorts of student results data from the largest national education board in India, the Central Board of Secondary Education (CBSE), to examine peer effects of students belonging to Scheduled castes (SC) and Scheduled tribes (ST), the most marginalized communities in the country, on test scores in national level standardized examination for students in the highest grade (class XII) in school.

While the reservation system mandates 22.5% seats for SC and ST persons in political representation, government jobs and higher education, the Right to Free and Compulsory Education Act (2009), commonly known as RTE, went one step further and mandated a minimum of 25% reserved seats for children of economically

weaker and socially disadvantaged groups in all *primary* unaided private schools. Though this policy is not applicable in our setting where we look at test scores of students in the highest grade in school, our motivation comes from the general perception and apprehension that the presence of disadvantaged caste students in a student's peer group will negatively affect her own performance (*Frontline* (15 July, 2011); *Hindustan Times* (New Delhi, 19 September, 2012); *News 18* (Saharanpur, 14 April, 2018); *The Print* (17 February, 2019)).

Identification of causal peer effects is tricky because students can select into schools endogenously. To address this, we identify casual effects using the variation in the peer composition of adjacent cohorts within a school (Hoxby 2000). By including school fixed effects in our empirical specification, we are able to eliminate the endogeneity bias stemming from self selection into schools in a given cohort. In addition, we include school-specific linear time trends to control for any time-varying unobservables at the school level. Finally, we also include a student's past scores as proxies for ability and past inputs into the education production function (Hanushek 1979).

Our results show that the above stated apprehensions are baseless. After controlling for a string of student socioeconomic characteristics, past scores, school and cohort fixed effects and school-specific linear time trends, there is no statistically significant effect of the cohort-to-cohort variation in the share of SC/STs in a student's peer group on her test scores in the national level standardized class XII board examination. These results are precisely estimated, so that we can reject modest sized effects between 0.12σ and 0.14σ. We conduct a host of robustness checks and find that the null effects hold separately and are estimated precisely for all the caste groups, both genders, all income quartiles, private and public schools and for stu-

dents with different stream choices in higher secondary school. We also show that the null effects do not mask heterogeneous effects by ability: the results are statistically insignificant across the ability distributions of both students and peers.

Chapter 2

Whose education matters? An analysis of inter caste marriages in India

2.1 Introduction

Ethnic endogamy as a practice to entrench clan, community or tribal boundaries has been around for centuries (Davis 1941; Bisin and Verdier 2000). In the Indian context too, endogamy is central to the institution of caste.[1] Indian castes are largely endogamous groups and violations of caste endogamy are often punished by social ostracism (Chowdhry 1997; Kaur 2010; Bidner and Eswaran 2015). It is also one of the most resilient caste based practices till date. The rate of inter caste marriages,

[1] A huge body of literature has been developed to understand the origin, nature and contemporary aspects of the caste system in India. While it is too vast to be summarized here, see Srinivas (1962), Beteille (1971) and Dumont (1980), for some seminal works in this area. For excellent surveys of the literature, see, for example, Vaid (2014), Munshi (2017), Mosse (2018).

even as recent as in 2011, was as low as 5.82% and there has been no upward time trend over the past four decades[2]. In this chapter we study the relationship of caste endogamy with education, taking into account the nature of the Indian marriage market where marriages arranged by parents and close relatives is largely the norm.

Two aspects of the institution of caste highlight the importance of inter caste marriages. First, the caste system has been shown to be discriminatory (Shah 1985; Thorat and Newman 2007), and detrimental to democracy (Jeffrey 2002; Munshi 2017), social mobility (Munshi and Rosenzweig 2006; Munshi 2017), trade (Anderson 2011) and environment (Gadgil and Rao 1994). Second, caste endogamy being the pillar of caste system (Bidner and Eswaran 2015), inter caste marriages can directly weaken the foundations of caste system. In addition, though not directly for caste, there exists evidence of positive impact of intermarriages. For example, intermarriages between natives and different immigrant ethnicities are associated with higher immigrant wages (Meng and Gregory 2005) and higher female labour supply (Gevrek et al. 2013; Wong 2014) in the context of Australia, Canada and the USA, respectively. Kalmijn (2010) shows strong evidence that interracial marriages have integrative effects on the offsprings for the case of the Netherlands. Positive effects of inter-ethnic marriages have also been shown on the social, cultural and economic integration of the children in England, Germany, the Netherlands and Sweden by Kalmijn (2015) and in two American cities by Stephan and Stephan (1991). Finally, extreme manifestation of endogamy in the form of consanguineous marriages may even be inefficient from the perspective of democracy and it may promote corruption and nepotism (Luke and Munshi 2006; Schulz 2019; Carl 2017; Akbari et al. 2019, 2020)[3].

[2]Authors' calculations from the data set used for the study, the second round of the Indian Human Development Survey.

[3]The literature on exogamy also discusses a few negative aspects associated with them. Two

A number of factors may influence the marriage choice of an individual. Since we are interested in looking at inter caste marriages in the particular context of weakening the institution of caste, we explore how education is associated with the probability of an inter caste marriage. Dr. Bhim Rao Ambedkar, the chief architect of the Constitution of India and one of the tallest leaders of the disadvantaged castes, was of the view that education could free the marginalized sections of the society and set them on a path of upward mobility (Velaskar 2012; Moon and Narke 2014a,b; Zene 2018). This spirit is incorporated as the primary focus of all education policies of India (National Policy on Education 1968, 1986; Right of Children to Free and Compulsory Education Act 2009; Joshee 2008; Mander and Prasad 2014). In addition, Dr. Ambedkar proposed that inter caste marriages will directly weaken the caste system (Ambedkar 1936). The same idea has been propounded by the Indian judiciary as well as policymakers(https://www.timesnownews.com/mirror-now/in-focus/article/inter-caste-marriages-should-be-encouraged-for-uprooting-caste-system-madras-high-court/440195; Ambedkar Scheme for Social Integration through Inter-Caste Marriages 2016).

In this chapter, we aim to establish a link between education and inter caste marriages since education can not only mitigate deeply held prejudices, educational

major themes in this literature are the associations between exogamy and marriage dissolution rates, and the outcomes of the off-spring of exogamy. While the second theme has mostly found no negative and a few positive effects of exogamy on the children, there is stronger evidence in the literature that exogamy is correlated with lower family stability. For example, Kalmijn et al. (2005) show that heterogamous marriages are more likely to end in divorce than homogamous marriages in the case of inter-religious and inter-national marriages in the Netherlands. Similarly, Milewski and Kulu (2014) and Dribe and Lundh (2012) find that divorce rates are higher in out-marriages with greater social and cultural distances between the spouses in case of Germany and Sweden, respectively. The literature has also looked at labor market outcomes and time use outcomes of couples in out-marriages. Basu (2015) finds negative and statistically significant effects of intermarriage on the wages of Asian women in the USA. Grossbard et al. (2014) uses the American Time Use Survey Data and finds that, in general, blacks do more housework in inter-racial unions than in all black unions.

institutes can also serve as platforms for social mingling, especially since inclusive education has been a primary focus of the Indian education policy (National Policy on Education 1968, 1986; Right of Children to Free and Compulsory Education Act 2009; Joshee 2008; Mander and Prasad 2014).

A large part of the literature on out-marriages focuses on its relationship with the education of individuals and the evidence is mixed depending on the context and the study sample. Qian (1997) and Fryer (2007) find a positive relationship between educational attainment and the likelihood of an interracial marriage in the US. While Qian and Lichter (2001) find this relationship to be positive for Latinos, Hwang et al. (1995) find, in contrast, that Asian women with lower levels of education are more likely to out-marry racially. Gullickson (2006), on the other hand, does not find any consistent relationship between education and the likelihood of interracial marriages for whites.

Studies on exogamy in South Asia have been relatively scarce and primarily based on localized samples (Dugar et al. 2012; Banerjee et al. 2013; Allendorf and Thornton 2015; Ahuja and Ostermann 2016). To the best of our knowledge, we are the first to make a systematic attempt at understanding the relationship of education with inter caste marriages in India using a nationally representative data set. But, at the outset, we recognize that we have to pay due attention to the fact that marriage markets in India work very differently as compared to the Western countries (Banerjee et al. 2013). A majority of marriages are arranged by the parents, and the spouses barely know each other before marriage. In our data set (second round of the Indian Human Development Survey, IHDS-II), 73% of marriages were reported to have been arranged by parents and almost 70% of the women said that they met their husbands only on the day of their wedding/*gauna*[4]. This pattern, quite

[4] *Gauna* is a ceremony conducted after several years of a child marriage when the bride moves

surprisingly, holds for the inter caste marriages as well: close to 63% of those who said they were in an inter caste marriage reported their marriages to be arranged by parents. In fact recent studies using the IHDS have shown that even over time, the movement has not been towards "Western-style marriage, in which young people choose their own spouses" (Allendorf and Pandian 2016). The shift is rather towards increased say of women within the purview of "arranged marriages"[5] (Banerji et al. 2013; Allendorf and Pandian 2016).

The wide prevalence of the arranged marriage institution in the Indian marriage markets strongly suggests that any analysis of marriages in India must consider parental attributes along with individual ones. To justify this approach, we first explore whether education levels of the spouses themselves have any predictive power on the likelihood of inter caste marriages. We find that, contrary to the findings in the existing literature on out-marriages in the West, especially in the USA, the education levels of the individuals themselves do not have any association with the probability of inter caste marriages. The result is very robust to the inclusion of a whole range of controls and fixed effects, and to variations in the sample.

To examine our null results, we attempt at disentangling two potentially opposing effects of education identified by Furtado (2012). The first is the 'cultural adaptability effect' through which education makes members of different groups more aware of and adaptable to the culture of each other and hence, may increase the incidence of intermarriage. The second one, the 'assortative matching effect'[6], however, may work in either direction. In a group with average education level below (above) the

from her natal home to her husband's family.

[5]The term "arranged marriage" is used to refer to a marriage where parents or other relatives play the main role in selecting a spouse for their offspring, often keeping social attributes like caste and economic status of the family in view (Banerji et al. 2013).

[6]The term assortative matching refers to a positive correlation between the attributes of the husband and the wife. In our case, for example, the attribute is education.

average education level of the relevant population, a more educated individual will marry out (marry in) and education will have a positive (negative) effect on exogamy for that group. The net effect can go in either direction and one may observe a positive, a negative or no relationship between education and exogamy depending on a particular group's characteristics. We adapt the methodology suggested by Furtado (2012) to the Indian context and our original findings are reaffirmed. None of the channels have any statistically significant association with the probability of inter caste marriages in India.

Our null results can mask important heterogeneity across caste groups. According to the Status Exchange theory (Davis 1941; Merton 1941; Kalmijn 1998; Gullickson 2006; Fu and Heaton 2008)[7], in an inter caste marriage, the upper caste individual will typically be able to exchange her/his caste status for a higher level of education of a spouse from a lower caste as compared to the level of education of the spouse she/he would be matched to in an *intra* caste marriage. As a result, the marginal effect of an increase in education will be higher for a lower caste individual compared to a higher caste individual in the inter caste marriage market because education can be exchanged for caste status. We check for such heterogenity but find, very similar to Banerjee et al. (2013), no evidence of status exchange taking place.

Having established the irrelevance of the spouses' own education, we next explore whether parental education is associated with the likelihood of an inter caste marriage. We add the education levels of the parents of both the spouses to our set of explanatory variables. Here we find that the level of education of the husband's

[7]The status exchange theory broadly postulates that an intermarriage, especially between two groups which are unequally ranked in the social hierarchy, often involves an exchange of characteristics between the two parties such that both stand to gain from the union. Typically one party exchanges its social status for some other trait of the spouse, like beauty or education. Thus, more educated blacks would marry less educated whites because they would gain from a higher social status of their spouse (Gullickson 2006).

mother has a positive and statistically significant association with the likelihood of an inter caste marriage. One standard deviation increase in the years of education of the husband's mother is associated with a 10.16% increase in the probability of inter caste marriage over the sample mean. The result is very robust to variations in the sample, to the addition of a number of controls as well as fixed effects, to alternate model specification and to omitted variable bias (Oster 2019). However, this part of the result is nuanced in the sense that among the parents on both sides, only the education of the husband's mother has a predictive power on the likelihood of inter caste marriage. Given our dataset, we are unable to empirically establish a precise channel for this finding. However, we posit some potential channels based on theoretical arguments from the existing literature and provide some suggestive evidence for our proposed mechanism.

The rest of the chapter is organized as follows. In section 2.2, we describe the data. The descriptive analysis in section 2.3 prepares the contextual background and provides the descriptive statistics. Sections 2.4, 2.5 and 2.6 detail the regression analysis, specifying the empirical strategy and discussing the main results and robustness checks, respectively. Section 2.7 gives a brief discussion of the possible channels behind the results and section 2.8 concludes.

2.2 Data

We use data from the latest round of the Indian Human Development Survey (IHDS II). The IHDS is a nationally representative household panel survey conducted in 384 districts, composed of 1420 villages and 1042 urban neighborhoods across all states and union territories of India. The second round of the survey, IHDS-II,

was conducted in 2011-12.[8] The survey has detailed socio-economic and human development related questions for a household as a whole, for young children in the household and for one ever married woman in the age group of 15-49 years in each household called the 'eligible woman'. We combine data from two schedules of the survey. The household schedule contains detailed questions about various socio-economic characteristics of the household. In the eligible woman's schedule, one eligible woman was interviewed regarding health, education, fertility, family planning, marriage and gender relations in the household and the community.[9]

Even though caste and various caste based practices are common in India, there has been little systematic attempt so far to collect data on these aspects in a nationally representative survey. IHDS, for the first time, asks questions that help us explore along this direction. Our outcome variable, whether a marriage is an inter caste marriage, is defined using the following question in the eligible woman's schedule: "Is your husband's family the same caste as your natal family?" The dependent variable "ICmarriage" takes value 1 if the answer to this question is "No". This question accurately reflects whether a marriage is inter caste or not since the marriage is recognized by the responding woman as inter caste and hence is "ultimately closer to the lived reality of an inter-caste marriage".[10] It is important to point out here

[8]IHDS II re-interviewed 83% of the original as well as split households residing within the village which were interviewed in IHDS-I, and an additional sample of 2134 households.

[9]In the households where the eligible woman from the first round of the survey died between the survey waves or was no more in the eligible age group, a new eligible woman was interviewed, along with the old one, if present. Thus there can be a maximum of 2 eligible women in each household. In households with more than one potential eligible woman, one was selected using a standard random number procedure in IHDS-I (Desai et al. 2009).

[10]According to The Hindu (New Delhi, 13 November 2014) (Rukmini 2014), the IHDS said that "... what female respondents interpreted as a "different caste" is likely to have been subjective, but ultimately closer to the lived reality of an inter-caste marriage". In her interview to The Hindu, Sonalde Desai (Senior Fellow at NCAER and Professor of Sociology at the University of Maryland) who led the IHDS, said: "So the IHDS took a simple approach and asked women whether their natal family belongs to the same caste as their husband's family, allowing us to bypass the complex issue of defining what caste means and get subjective perceptions from our respondents".

that although in the English version the word caste is used, the question actually uses the word "*jati*" in Hindi (and its equivalents in all the other eleven languages in which the survey was administered), and not caste. This takes care of the fact that the finer *jati* level is relevant for marriages in India and not the caste level, which is often synonymous with the broad administrative categories in India.

Our main independent variables of interest are the years of education of the spouses and their respective parents. They range from 0 (illiterate) to 16 (above graduate) years. Our set of control variables include the caste and the urban or rural location (according to Census 2011) of the husband's household at the time of the survey. We include assets (index created by IHDS) and annual per capita income (in INR) of the household at the time of the survey to proxy for the assets and income level of the household at the time of the marriage. Finally, we control for the age at marriage of the bride and the comparative economic status of the two families at the time of their marriage.

We use three rounds of the Employment and Unemployment Survey of the National Sample Survey of India (NSS) conducted in 2004-05, 2009-10 and 2011-12 to construct average and caste-wise average years of education of females in the marriageable age group (12 to 35 years) for each district at the time of marriage.[11,12] We also calculate the proportion of population belonging to the same caste as that of a husband in our sample in his district of residence using these NSS data sets. These variables are used to separate the opposing effects of education, namely, cultural adaptability and assortative matching effects.

[11]The marriageable age group is constructed by looking at the distribution of age at marriage of the eligible women in the IHDS sample where 96.8% of women report their age at marriage to be from 12 to 35 years.

[12]The nature of the NSS data and the fact that inter-district migration due to marriage is very low in India (Desai and Andrist 2010; Stopnitzky 2012) helps us in constructing these variables at the time of marriage and not just for the NSS survey years.

2.3 Descriptive Analysis

Our specific aim in this chapter is to look at the relationship between inter caste marriages and education. We set the stage by looking at a broad range of descriptive statistics to get a better idea about the existing trends and dynamics of the marriage market in India in general and inter caste marriages in particular.

Figure 2.1 plots the rate of inter caste marriages by the year of marriage.[13] Even in the face of industrialization and urbanization in India, an upward trend is not visible over the last four decades: the rate of inter caste marriages has hovered around 5% since 1970 to 2012.[14] The average for 2000-2012 is marginally higher than 1971-80 and 1981-90, but is not statistically different from the decade of 1990-2000.

In Table 2.1 we look at the distribution of inter caste marriages by various characteristics of the husbands' households. The first panel shows that *Brahmins* have the highest rate of out-of-caste marriages, followed by Other Forward castes (OFC), while Other Backwards Classes (OBC) and Scheduled castes (SC) have the lowest rate.[15] However, the rate of exogamy for *Brahmins* is not statistically different from any other caste categories. The only significant differences are between the rates of OFCs and OBCs, and OFCs and SCs.[16]

[13]In IHDS II the year of marriage variable has 30.66% missing values. We, therefore, construct our own variable for the year of marriage using the year of birth of the eligible woman respondent and her age at marriage.

[14]The Modernization theory in Sociology explains the process of transition of a nation from a traditional political structure to a democratic one via causal chains of industrialization, urbanization, education and so on (Przeworski and Limongi 1997). One of the predictions of the Modernization theory is that with the advent of industrialization and urbanization, various non-Western family behaviours will converge towards the Western nuclear family model. As a result, there will be a decline in arranged marriages, which "... likely signals declines in the importance of ethnicity/caste, religion ..." (Allendorf and Pandian 2016).

[15]Refer to the Appendix to this chapter for a description of the social and administrative categorizations of the caste system in India.

[16]A reported inter caste marriage may not necessarily involve two broad administrative caste categories.

The second panel of Table 2.1 shows that the rates of inter caste marriages are not statistically significantly different between urban and rural households. A finer division tells us that within the urban sector, it is the metropolitan urban areas that have the lowest rate, while other urban areas have a higher rate (3.84% and 5.41% respectively). Within the rural sector, developed villages have a higher rate, while less developed villages have a lower rate of inter caste marriages (5.72% and 4.86% respectively). Thus more urbanized areas do not necessarily have a higher rate of out-marriages in India.

The next two panels of Table 2.1 show the rate of inter caste marriages by asset and annual per capita income quartiles of the households respectively. In both cases the rate goes down as we move up the distribution (poorest to the richest): the rate of inter caste marriages is significantly higher in the first quartile than that in the fourth quartile. The last panel of Table 2.1 shows that no difference is observed in the rate of inter caste marriages irrespective of whether the husband's family had the same, better or worse status than the wife's family at the time of their marriage. The observations so far make it clear that caste endogamy is much more pervasive than expected in the face of economic development and expansion of market forces.

In Table 2.2 we look at the decision making process at the time of marriage. The second column of Table 2.2 reports the percentages among all marriages while the third column reports that among inter caste marriages only. Among all marriages, a striking 73% of women say that parents (or other relatives) chose their husbands, and in fact almost 70% of them met their husbands only on the day of their wedding/*gauna*. Only a quarter of the women had met their husbands or had seen their photos before marriage; even fewer had talked to their husbands before getting married to them (third panel of Table 2.2).

Even among the subset of only inter caste marriages, almost 63% of them are arranged by parents/other relatives only. Interestingly, even here an overwhelming 98.07% of couples lived with their parents immediately after marriage. Thus, when a marriage takes place, inter caste or not, the parents have the primary say in a majority of the cases. This observation lends reasonable amount of support to the idea that the effect of parental attributes should be central in any analysis of marriages in India.

Finally we turn to our main attribute of interest, namely education. Figures 2.2 and 2.3 plot the rate of inter caste marriages for different educational categories of the wife and the husband, and wife's mother, wife's father, husband's mother and husband's father respectively.[17] Figure 2.2 shows that this rate is not statistically significantly different among the different educational categories of the spouses themselves.

From Figure 2.3 it can be observed that the rate of inter caste marriages does not vary by the educational categories of the fathers of the spouses(The mean differences between any pair of educational categories of the fathers are statistically insignificant in general.). However, the rate of inter caste marriages appears to be significantly higher at higher educational categories of the mothers of the spouses(The mean differences are statistically significant and positive for a number of pairs of educational categories.). This corroborates well with the earlier observation that parental attributes should be important in the analysis of marriages in India where the institution of arranged marriages plays a dominant role. In what follows, we further explore along these directions in a regression analysis of the relationship between

[17]These categories are constructed by dividing the years of education into five bins: Illiterate (0 years), Up to Primary (1 to 5 years), Up to Secondary (6 to 10 years), Up to Bachelors (11 to 15 years) and Above Bachelors (more than 15 years).

inter caste marriages and education.

2.4 Empirical Framework

Our observations in the previous section suggest that marriages in India are arranged primarily by parents with minimal say of the individuals themselves. Thus we must pay due attention to parental education along with the education of individual spouses.

We, therefore, proceed in two steps. First, we explore whether education levels of the spouses themselves can predict the occurrence of inter caste marriages. Considering a married couple as our unit of observation, we run the following regression:

$$\begin{aligned} ICmarriage_{id} = \alpha + \beta_1.husband's\ education_{id} + \beta_2.wife's\ education_{id} \\ + \theta.X_{id} + \delta_d + \tau_t + \varepsilon_{id}. \end{aligned} \tag{2.1}$$

Here $ICmarriage_{id}$ is a binary variable which takes value 1 if a couple i in district d is in an inter caste marriage and 0 if in an intra caste one. Our primary independent variables of interest are the education variables: *husband's education*$_{id}$ denotes the years of education attained by the husband and *wife's education*$_{id}$ is that attained by the wife.

In the next step we add the years of education of the parents of both the spouses to the set of explanatory variables considered in equation (2.1):

$$
\begin{aligned}
ICmarriage_{id} = \alpha &+ \beta_1.husband's\ education_{id} + \beta_2.wife's\ education_{id} \\
&+ \gamma_1.husband's\ mother's\ education_{id} + \gamma_2.husband's\ father's\ education_{id} \\
&+ \gamma_3.wife's\ mother\ education_{id} + \gamma_4.wife's\ father's\ education_{id} \\
&+ \theta.X_{id} + \delta_d + \tau_t + \varepsilon_{id}.
\end{aligned} \tag{2.2}
$$

Similar to equation (2.1), *husband's mother's education*$_{id}$, *husband's father's education*$_{id}$, *wife's mother's education*$_{id}$ *and wife's father's education*$_{id}$ are the completed years of education of the husband's parents and wife's parents respectively.

In both equations (2.1) and (2.2), X_{id} is a vector of couple and household level control variables, namely, administrative caste category of the husband's household (Brahmins, OFC, OBC or SC), age at marriage of the wife and dummies for the comparative economic status of the two families at the time of the marriage. It also includes the per capita income and the assets index of the household and its location (rural or urban).

Marriages in India occur overwhelmingly within the district (Desai and Andrist 2010; Stopnitzky 2012). Therefore, we include district fixed effects, δ_d, to control for any time invariant unobserved factors at the level of a district. We also include year of marriage fixed effects, τ_t, to control for all unobservables across districts in the year a couple got married.

In our data set, households belonging to all religions report their castes. However, the caste system was originally a Hinduism phenomenon. To incorporate both these observations, the sample for our main analysis consists of only those households which have stated their religion as Hinduism, Buddhism, Jainism or Sikhism. Our choice is driven by the fact that all these religions come under the Hindu Marriage

Act of the Constitution of India. We also exclude scheduled tribes (STs) from our main sample mainly because even though a significant number of tribals report their religion as Hinduism, "there is sufficient heterogeneity and distinctiveness within tribal communities that they cannot be considered a part of the *varna* system". (Deshpande 2011)[18] For our analysis we consider the 20 major states of India.[19] Our final sample consists of 25,070 couples of which 1079 couples have inter caste marriages. Standard errors are clustered at the Primary Sampling Unit (PSU) level. Table 2.3 provides the summary statistics for all the variables used in the regressions. All calculations use the survey weight of the eligible woman.

2.5 Results

2.5.1 Inter caste marriages and own education

Table 2.4 reports our first set of results. The first two columns report results from the estimation of equation (2.1). In column 1, the regression coefficients from the parsimonious specification with only caste controls and the education levels of the spouses show that the education of neither the husband nor the wife is associated with the likelihood of an inter caste marriage. In column 2, we add the full set of our control variables. The addition of these controls has no effect on the coefficients of the spouses' own education – they remain statistically insignificant. This result stands in sharp contrast to the findings in the existing literature on out-marriages in the Western countries where individual's own education shows up as a predictor

[18]Refer to the Appendix to this chapter for a description of the social and administrative categorizations of the caste system in India.

[19]This list includes the following states: Himachal Pradesh, Punjab, Uttarakhand, Haryana, Delhi, Rajasthan, Uttar Pradesh, Bihar, Assam, West Bengal, Jharkhand, Orissa, Chhattisgarh, Madhya Pradesh, Gujarat, Maharashtra, Andhra Pradesh, Karnataka, Kerala and Tamil Nadu. We exclude the states of North-East, Goa and Jammu and Kashmir.

of one's marriage being within or outside one's race or ethnicity.

To investigate our null results, in the next three columns, we test whether any of the mechanisms of the effect of education as described in Furtado (2012) come out to be statistically significant. We adapt the model suggested by Furtado (2012) to the Indian context:

$$\begin{aligned} ICmarriage_{icd} = \kappa &+ \lambda.husband's\ \ education_{icd} + \pi_1.(avg\ \ FemEdu_{cd} - avg\ \ FemEdu_d) \\ &+ \pi_2.husband's\ \ education_{icd}.(avg\ \ FemEdu_{cd} - avg\ \ FemEdu_d) \\ &+ \mu_1.population\ \ proportion_{cd} + \mu_2.population\ \ proportion_{cd}^2 \\ &+ \sigma.X_{id} + \Psi_s + \tau_t + \xi_{icd}. \end{aligned} \tag{2.3}$$

The dependent variable is a dummy which takes value 1 if husband i of caste c in district d is in an inter caste marriage.[20] The first term on the RHS, husband's years of education, captures the cultural adaptability effect of education. If the analysis of Furtado (2012) holds for our sample, this coefficient should be positive: an increase in education makes an individual more accepting and adaptable to the culture of other castes. The next term, $avg\ \ FemEdu_{cd}$ is the average education level of females in the marriageable age group (12 to 35 years) in the husband's caste in his district and $avg\ \ FemEdu_d$ is the average education level of *all* females in the marriageable age group in his district.[21]

The coefficient π_2 measures the assortative matching effect of education, which is captured by the interaction term of husband's years of education with the deviation of average education of females within his caste in the district from the average

[20]Since we do not know the caste of the wife in a couple, our sample consists of only husbands for this set of regressions.

[21]Both the variables, $avg\ \ FemEdu_{cd}$ and $avg\ \ FemEdu_d$, have been calculated at the district level and correspond to the relevant couple's year of marriage.

female education in the entire district.[22] The expected sign of π_2 is negative if the assortative matching effect of education is at work. A man with a higher level of education is more likely to find a higher educated woman from his own caste if the average education level of the women of his caste is higher than the district average.

We also include the proportion of female population in the marriageable age group of husband's caste in his district, $population\ proportion_{cd}$, which captures the enclave effect: likelihood that the individual will encounter a potential spouse of the same caste in his relevant region of search, which we take to be the district based on the literature (Desai and Andrist 2010; Stopnitzky 2012).[23]

Column 3 of Table 2.4 contains results from a regression similar to that of the second column, but uses only the husband's education variable (and replaces district fixed effects with state fixed effects) to make it comparable to the regressions in the next two columns. This coefficient, capturing the cultural adaptability effect, is still statistically insignificant. In column 4, we add the assortative matching term. The estimated coefficient of this variable is statistically insignificant and it also does not affect the coefficient of husband's own education.[24] Finally in column 5, we add the enclave effect term. The addition of this control and its square term too have no effect on insignificance of the coefficient of the husband's education variable. The coefficients on the variables themselves are also statistically insignificant.

Thus, even after we explicitly take into account the potential channels, as analyzed in Furtado (2012), through which own education might have an effect, we find that neither of these channels predict the likelihood of an inter caste marriage.

[22]The coefficient π_1 captures the main effect of this deviation.

[23]For this set of regressions, we include state fixed effects, Ψ_s, instead of district fixed effects because our regressors are district level variables.

[24]We also calculate the education difference term by excluding husband's own caste females from the district average and use this variable in our regressions. All our results remain the same.

As noted in the introduction, our null results might mask important heterogenity across caste groups. According to the status exchange theory, one party exchanges its social status for some other trait of the spouse, like beauty or education. Hence we might have a positive association between education and exogamy for some caste groups and negative for some other groups leading to a net null association between education and exogamy for all caste groups taken together. We run another set of regressions to check this but we find no evidence of status exchange[25]. Our result is very similar to Banerjee et al. (2013) who also find almost non-existent preference for "marrying up" or exchanging other attributes for caste status.

2.5.2 Inter caste marriages and parental education

Now we move on to add the education level of the parents of both the spouses to our set of explanatory variables. For the sake of comparison, column 1 in Table 2.5 reproduces the column 1 of Table 2.4. Column 2 reports results from the estimation of equation (2.2) where we add the education levels of the parents of the spouses. We find that the education of the husband's mother has a positive and statistically significant association with the probability of an inter caste marriage. A one-year increase in education of the husband's mother increases the probability of an inter caste marriage by 0.18 percentage points. The results in both the columns 1 and 2 are consistent with our descriptive analysis where we observed that parents have the major say in any marriage in India and individuals themselves have a very little role to play.

In columns 3 and 4 we successively add controls to the base specification[26]. The

[25]Refer to the Appendix to this chapter for a detailed discussion of the status exchange theory, our empirical specification and the regression results.

[26]In column 3 we add the age at marriage of the wife and dummies indicating whether the economic status of the wife's natal family was better, same or worse than that of the husband's

addition of these variables has little effect on the coefficient of the husband's mother's education.

We conclude this section with the key finding that husband's mother's education positively predicts the likelihood of an inter caste marriage and that it is robust to the inclusion of a number of controls and fixed effects. A one standard deviation increase in husband's mother's years of education leads to a 10.16% increase (over the sample mean) in the probability of the couple's marriage being an inter caste one. To put this in perspective, we compare the effect size of education on exogamy between the Indian and US data. Based on calculations made from Furtado (2012), we find that a one standard deviation increase in education of the husbands in her sample leads to only a 7.08% increase in his likelihood of inter ethnic marriage. A similar calculation shows that a one standard deviation increase in the husband's mother's years of education (in our data) explains 46.85% of the increase in the rate of inter caste marriages from 1970 to 2012.

2.6 Robustness checks

We report four robustness checks in Table 2.6. In column 1, we remove the women who continued their education post marriage as this could potentially contaminate the results since these women will actually have a lower amount of education at the time of their marriage as compared to what is measured by the data. All our results are qualitatively the same even for this sample.

It is plausible that if women had a greater say in their marriages, it may bias the coefficient on the education of the husband's mother upwards. A greater decision

family at the time of the marriage. In column 4, we add current income and assets of the household, and whether the household was located in an urban or rural area.

making power of the brides in their marriages may be positively correlated with both higher education of her husband's mother as well as with the probability of an inter caste marriage. Therefore, in column 2, we look at the sample of only parents-arranged marriages, or simply arranged marriages as they are commonly known. We define arranged marriages as marriages in which the eligible woman's response to the question "Who chose your husband" was either "Parents/other relative alone" or "Others". It can be seen that even here own education of the spouses has no association but the education of the husband's mother has a positive and statistically significant association with the probability of an inter caste marriage.

In the third column of Table 2.6, we add another set of fixed effects to our controls – the interaction of district and year of marriage fixed effects, to control for any unobservables at the level of a particular district-year. The coefficient of husband's mother's education is still positive and statistically significant as can be seen from column 3. Also, spouses' own education does not show any association.

Since our dependent variable is binary, we report the estimation results from a logistic regression in the final column of Table 2.6. As can be seen, all our results go through with the logistic specification. The marginal effect of the husband's mother's years of education variable is 0.0025, while individual education coefficients are statistically insignificant as before.

We conduct another set of robustness checks to see if our results are robust to variations in the sample. We run our regressions for a Hindu-only sample, all-religions sample, all-castes (including STs) sample and all-states sample. We also use some other combinations: four main religions, main states, including STs; all religions, main states, including STs; and four main religions, all states, excluding STs. We find that our results are robust to all these sample variations. We report the first set

of these regressions in the Appendix to this chapter. The others are available upon request.

Although we do not claim any of our results to be causal, we still check if our results are being driven by unobservables. We examine the robustness of the result to omitted variable bias using the bound analysis methodology developed by Oster (2019). Here again we deduce that the coefficient of the husband's mother's education variable is not contaminated by omitted variables bias. The details of the methodology and our results can be found in the Appendix to this chapter.

Finally, since we test six simultaneous hypotheses (two education variables of the spouses and four of those of the parents), we also perform a series of multiple hypotheses corrections which control for Family-Wise Error Rate (FWER) as well as for False Discovery Rate (FDR). Our coefficient on the education of the husband's mother does not retain its statistical significance under these corrections. This is because given the structure of any of these tests, the lowest p value of the set is always corrected in a way similar to the Bonferroni correction (Farcomeni 2008) which is the most stringent correction (Abdi 2010; Fink et al. 2014; Streiner 2015). Since the lowest p value in our set is 0.033 (statistically significant at 5%), it is unable to retain significance under any of the available multiple hypotheses correction procedures.[27] While we present our results with this caveat, we nevertheless conclude this section with a reasonable confidence in the robustness of our results to variations in the sample, to the addition of a number of controls, to the addition of a number of fixed effects, to a change in the regression model and to correction for omitted variable bias.

[27]However, we would also like to point out that the idea of multiple hypothesis correction has its criticisms and the available methods might lead to too high rates of Type II error (Perneger 1998; Ruhm 2003; Nakagawa 2004; Kim et al. 2013).

2.7 Discussion

Although our results do not have a causal interpretation, they do point out some interesting features of the Indian marriage market. Our analysis of the relationship between education and the age-old practice of caste endogamy in India highlights the importance of recognizing the arranged marriages institution in Indian marriage markets. We first establish the interesting result that the education levels of the individual spouses themselves do not have any statistically significant association with the probability of their marriage being an inter caste one. We complete our analysis by establishing that the education level of the husband's mother has a positive, statistically significant and quite large association with the likelihood of an inter caste marriage. All of our results survive a battery of robustness checks.

The second part of our findings is nuanced in the following two ways. First, only the education of the husband's mother predicts inter caste marriage, but not that of his father. Second, education of the wife's parents are not associated with the likelihood of an inter caste marriage. In what follows we try to offer a plausible mechanism to explain our empirical findings regarding the heterogenity in the relationship between the parents' education and the probability of inter caste marriage with the caveat that we cannot offer any direct evidence because of a lack of data.

To understand the first result we put together three stylized facts. Firstly, a large body of literature finds evidence that a more educated woman has an increased decision making power in a household.[28] In our own dataset too, we find some support for this claim by looking at the responses to various questions under the "Gender

[28]See, for example, Thomas (1994); Beegle et al. (2001); Banerji (2008); Doss (2013); Banerji et al. (2013). Banerji (2008) and Banerji et al. (2013) use IHDS I to show that education is associated with greater autonomy in partner choice decision and it strongly improves the individual's involvement in parent arranged marriages.

Relations" section asked to the eligible women.[29] Interestingly, one of the questions directly asks about who has the most say in the decision to whom the respondent's children should marry. We find that education of the respondent woman is positively and statistically significantly associated with the probability that she has the most say in this decision.

Secondly, it is also well documented, especially in the context of developing countries, that a mother is more responsive to the needs of her child, as compared to the father. Provided with resources, a mother is more likely to utilize them in the best possible interest of her children. A father, on the other hand, is more likely to spend it on various adult consumption goods like tobacco and liquor.(See, for example, Thomas (1990); Haddad and Hoddinott (1995); Lundberg et al. (1997); Phipps and Burton (1998); Duflo (2000); Duflo and Udry (2004); Friedberg and Webb (2006).)

Finally, from our own analysis and from the literature cited in previous sections, we know that marriage decisions in India are taken by parents and other senior relatives and not by the prospective bride and groom.

Combining these three stylized facts we try to understand the first aspect of our finding as follows. Given that we are looking at marriages ex-post, the realized matches must be revealed preferred to be the optimal matches from all the potential matches available. An intra caste match could, then, be a constrained optimum if the father, driven by the prestige or reputation of the family and being less sensitive to the best outcome for the son, insists on the intra caste constraint. An inter caste marriage is more likely to occur when an educated mother can overcome this constraint and implement the best outcome for the son, empowered by her increased

[29]We find that the respondent woman's education is positively and statistically significantly associated with her likelihood of having the most say in seven out of the eight household decisions enquired in this section. The complete analysis can be found in the Appendix to this chapter.

bargaining and decision making authority in the family.

Consider next the second aspect of our finding that only the education of the husband's mother has a statistically significant association, but not that of the wife's parents. This asymmetry between the two families might arise from the fact that in any inter caste marriage the bride's family bears more stigma or costs than the groom's family. Some theoretical backing for this is provided by the analytical model in Bidner and Eswaran (2015) where stability of the endogamy equilibrium requires that the punishment for deviation from the equilibrium should be greater for a female and her family as compared to her male counterpart. While we could not find any empirical work on this asymmetry that arises in equilibrium, much of the anecdotal evidence involving "honour" killings in India validates our assertion[30]. Honour killing is killing someone in the name of family honour with the belief that the act will redeem the reputation of the family. It is often committed in cases where a couple marries against the wishes of the family, especially across caste lines. The fact that the crime is generally perpetrated by the bride's family, in which either or both of the spouses are killed, suggests that these families correctly expect to face the greater burden of the stigma of an inter caste marriage.

Our argument here is that education may not have enough mitigating effect on the stigma of an inter caste marriage for the bride's family which bears these costs disproportionately. Similar to the groom's father, the bride's father's education is not associated with the likelihood of inter caste marriage. However, unlike the case of the groom's mother, the education of bride's mother also has no association. This

[30]The Tribune, Chandigarh (03 July 2007): "*Honour killing rocks state, again*" (Manoj Babli honour killing Case); Times of India, New Delhi (20 November 2011): "*Parents held for 'honour' killing of 21-year-old Delhi University girl*"; The Indian Express, Ludhiana (09 May 2016): "*'Honour killing': Man kills daughter over relationship*"; Aljazeera (07 December 2016): "*India sees huge spike in 'honour' killings*".

difference may be due to the fact that unlike the groom's family, the bride's family bears a significant cost of an inter caste marriage. In other words, education works through giving more voice to the mother in the household to implement the best outcome for her child, if the stigma or social costs of an inter caste marriage is not too high.

2.8 Conclusion

We look at the relationship between education and the practice of caste endogamy, which is the defining and one of the most resilient features of the caste system in India. Using a nationally representative data set, the second round of the Indian Human Development Survey, we report novel and interesting findings. The rate of inter caste marriages in India is only 5.82% even in 2011, and there has been no secular increase in this rate over the previous four decades. In keeping with the existing literature, descriptive analysis of our data set shows that in the Indian marriage market families, rather than individuals, are the primary decision makers. An overwhelming 73% of marriages are arranged by parents, and spouses have very little contact with each other before marriage. Interestingly, this pattern holds true for inter caste marriages as well.

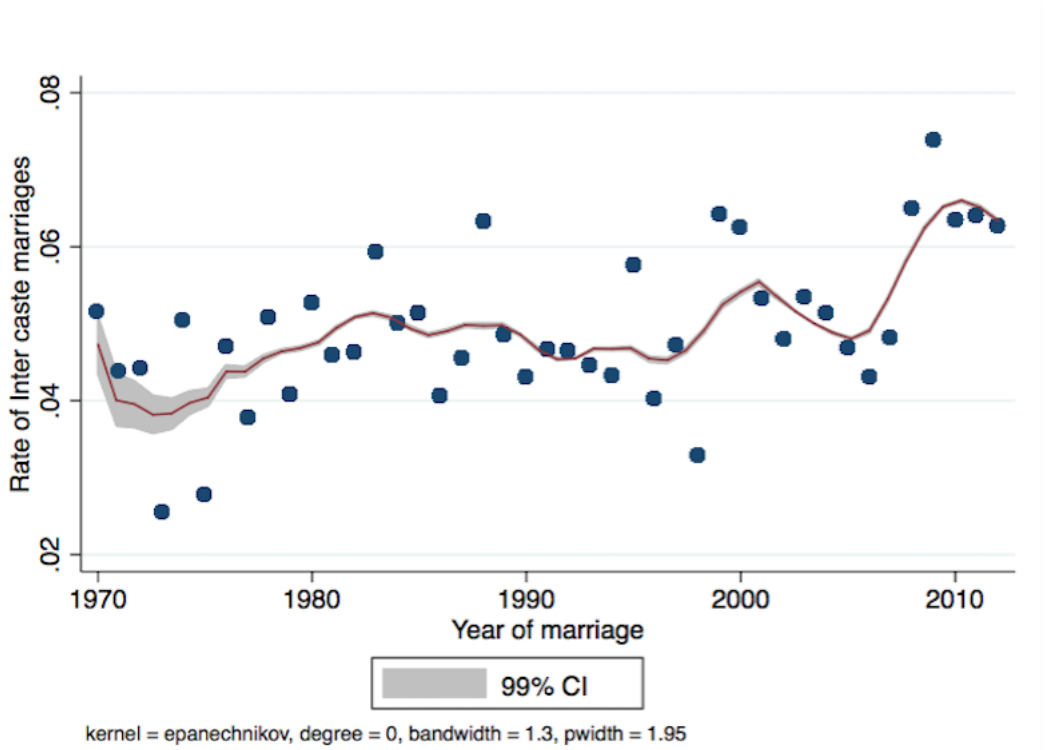
Our regression analysis brings out two important results. First, the education level of an individual does not predict the likelihood of his/her marriage being an inter caste one. In addition, we analyze if any of the possible channels suggested by Furtado (2012) is at work, but fail to find such evidence. We also see if there is any heterogenity in the relationship between education and exogamy across castes as suggested by status exchange theory, but do not find any such evidence. Second, complementing the observations from our descriptive analysis, we find that it is the

education of the husband's mother that has a positive and statistically significant association with the likelihood of an inter caste marriage. Both our results are robust to the inclusion of a host of control variables, a wide range of variations in the sample, *and* a varied set of fixed effects. Our results also stand the scrutiny of a logistic regression model as well as omitted variable bias using the bound analysis (Oster 2019). We posit that education works through giving more voice to the mother in the household to implement the best outcome for her child, if the stigma or cost of an inter caste marriage is not too large. Given that the bride's family disproportionately bears the stigma of an inter caste marriage, education of only the groom's mother has a positive association.

Our analysis highlights the importance of recognizing the institution of arranged marriage in any analysis of Indian marriage markets. Taken together, the two aspects of our result indicate that once the arranged marriage set up is recognized, one can easily understand the result that education has no effect on the decision of one's own marriage, rather it affects the marriage decision of one's offspring.

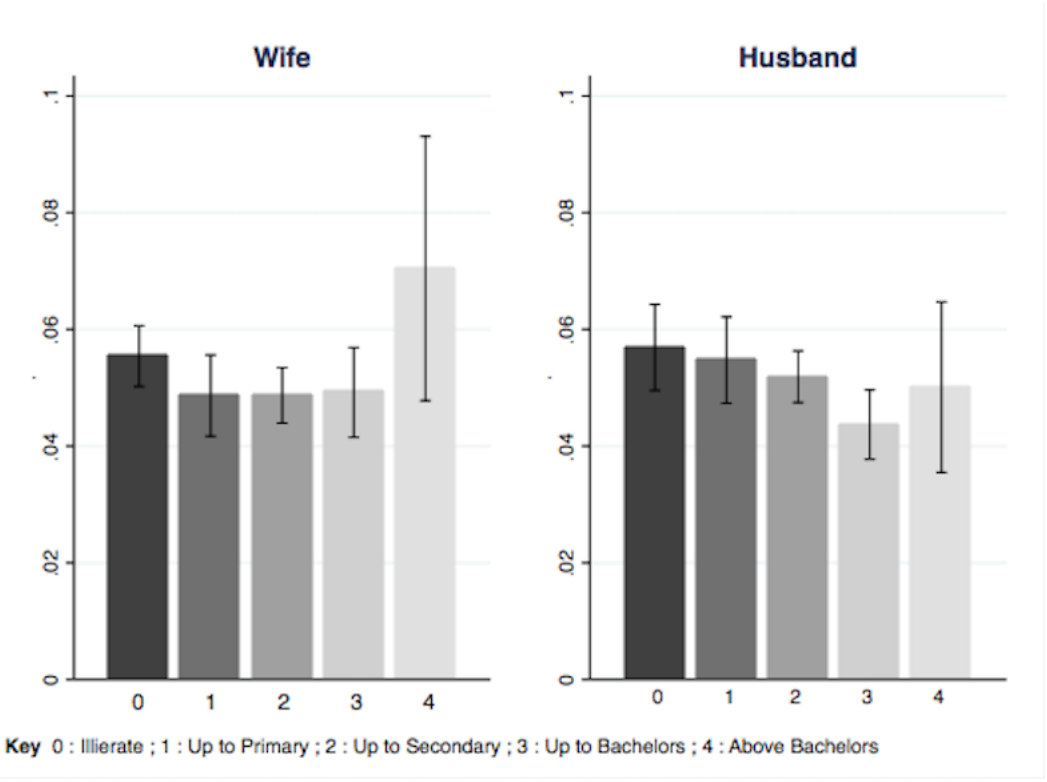
Figures and Tables for Chapter 2

Figure 2.1: Trend in the rate of inter caste marriages



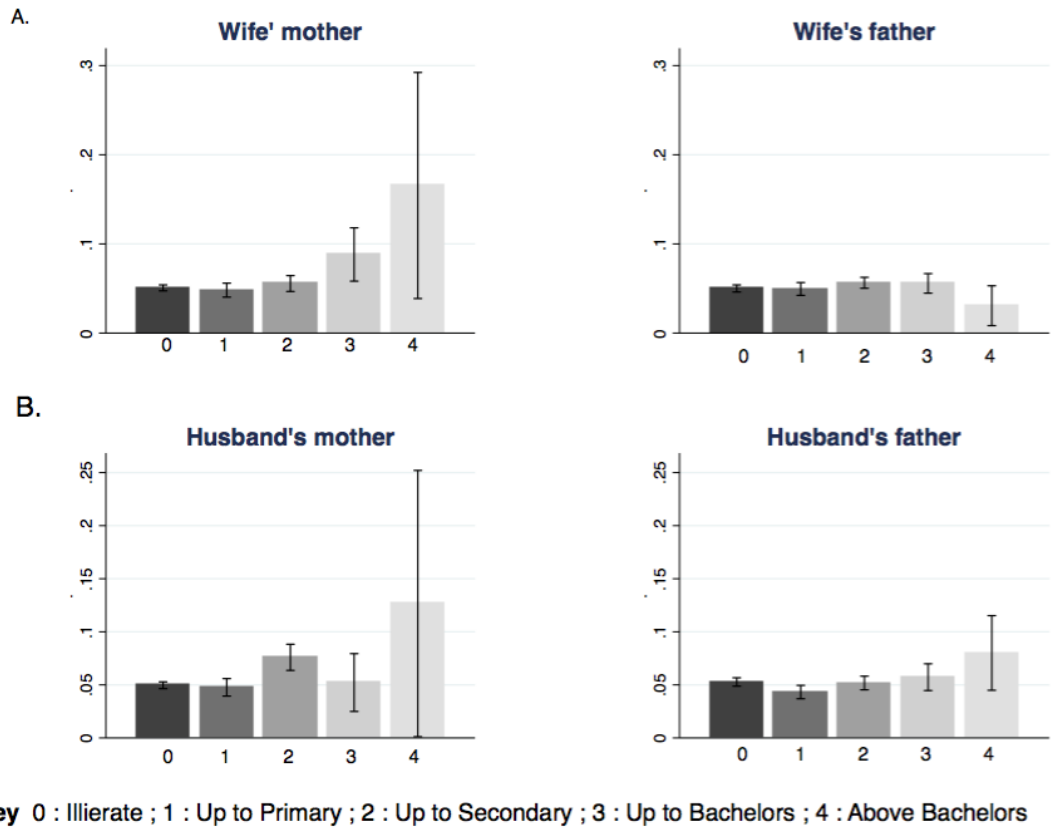
Note: The smooth line plots the local polynomial regression of the yearly rate of inter caste marriages on the year of marriage. Data source is IHDS II.

Figure 2.2: Rate of inter caste marriages and education of the spouses



Note: 95% confidence intervals indicated. Data source is IHDS II. The y axis stands for the rate of inter caste marriages. The left panel plots the rate of inter caste marriages by education of the wife while the right panel plots it by the education of the husband.

Figure 2.3: Rate of inter caste marriages and education of the parents



Note: 95% confidence intervals indicated. Data source is IHDS II. The y axis stands for the rate of inter caste marriages. Panel A plots the rate of inter caste marriages by the education of the wives' parents. Panel B plots the rate by the education of the husbands' parents.

Table 2.1: Rate of inter caste marriages by household characteristics

Caste	**Rate of Inter caste marriage**
Brahmins	6.30***
	(0.656)
Other Forward Castes	6.20***
	(0.341)
Other Backward Castes	4.80***
	(0.216)
Scheduled Castes	4.76***
	(0.269)
Type of Residence	
Urban	4.99***
	(0.246)
Rural	5.24***
	(0.184)
Asset quartiles	
First quartile (poorest)	5.89***
	(0.317)
Second quartile	5.48***
	(0.318)
Third quartile	5.01***
	(0.273)
Fourth quartile (richest)	4.01***
	(0.266)
Income quartiles	
First quartile (poorest)	5.08***
	(0.337)
Second quartile	5.58***
	(0.312)
Third quartile	4.07***
	(0.259)
Fourth quartile (richest)	4.89***
	(0.273)
Comparative Economic Status of wife's family (at the time of marriage)	
Same	4.98***
	(0.169)
Better	5.92***
	(0.387)
Worse	5.20***
	(0.480)

Note: Standard errors in parenthesis. Data source is IHDS II. The household here corresponds to the husband's household. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.2: Decision making at the time of marriage

Who chose the husband	**All marriages** (percent)	**Inter caste marriages** (percent)
Respondent herself	3.91***	15.01***
	(0.122)	(1.1.5)
Respondent and parents/other relative	22.70***	21.68***
	(0.286)	(1.33)
Parents/other relative alone	73.01***	62.83***
	(0.300)	(1.56)
Others	0.29***	0.49***
	(0.036)	(0.443)
Knew husband for how long before marriage		
On wedding/gauna day only	69.69***	66.5***
	(0.313)	(1.52)
Less than a month	13.33***	12.3***
	(0.232)	(1.06)
More than one month but less than one year	7.43***	5.82***
	(0.180)	(0.775)
More than one year	3.64***	11.7***
	(0.128)	(1.04)
Since childhood	5.46***	3.44***
	(0.155)	(0.588)
Met husband before marriage	23.43***	32.8***
	(0.287)	(1.52)
Saw photo of husband before marriage	26.72***	30.8***
	(0.301)	(1.49)
Talked to husband before marriage	15.64***	22.1***
	(0.246)	(1.34)
Chatted over email with husband before marriage	1.69***	3.45***
	(0.0856)	(0.591)
Living immediately after marriage		
With parents	99.2***	98.07***
	(0.062)	(0.445)
Alone	0.82***	1.93***
	(0.0615)	(0.443)

Note: Standard errors in parenthesis. Data source is IHDS II. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.3: Summary statistics

S.No	Variable	Mean	Standard Deviation
1	Inter caste marriage (binary variable)	0.0516	0.22
2	Wife's education (years)	5.51	4.95
3	Husband's education (years)	7.43	4.82
4	Husband's mother's education (years)	1.26	2.82
5	Husband's father's education (years)	3.33	4.42
6	Wife's mother's education (years)	1.63	3.18
7	Wife's father's education (years)	3.81	4.68
8	Age at marriage (Wife) (years)	17.61	3.55
9	Annual income per capita (INR)	25882.61	46471.64
10	Assets (Index)	15.76	6.46
11	Urban (binary variable)	0.3352	0.47

Table 2.4: Inter caste marriages and own education

	(1) ICmarriage	(2) ICmarriage	(3) ICmarriage	(4) ICmarriage	(5) ICmarriage
husband's education	-0.000351	-0.000102	-0.000429	-0.000454	-0.000413
	(0.000543)	(0.000520)	(0.000767)	(0.000782)	(0.000780)
wife's education	-0.000776	-0.000546			
	(0.000839)	(0.000814)			
(avg $FemEdu_{cd}-$ avg $FemEdu_d$)				-0.00184	-0.00176
				(0.00272)	(0.00276)
husband's education* (avg $FemEdu_{cd}-$ $avg\ FemEdu_d$)				-0.000130	-0.000158
				(0.000226)	(0.000231)
population proportion					0.0969
					(0.0765)
population proportion sq					-0.159
					(0.100)
Controls I		√	√	√	√
Controls II		√	√	√	√
Caste controls	√	√	√	√	√
Year of marriage FE	√	√	√	√	√
District FE	√	√			
State FE			√	√	√
N	22476	22469	22470	22027	22027
R^2	0.221	0.222	0.033	0.034	0.035

Note: Linear probability results are reported. Data sources are IHDS-II and Schedule 10 of NSS Rounds 61 (2004-05), 66 (2009-10) and 68 (2011-12). Outcome is a dummy variable which takes value 1 if the marriage is inter caste, 0 otherwise. The term *population proportion* is the proportion of population that belongs to the same caste as husband's caste and captures the potential enclave effect of education, *(avg $FemEdu_{cd}$ − avg $FemEdu_d$)* is the difference between the average education of females in the marriageable age in the husband's caste in his district and that of all females in the marriageable age in the husband's district and *husband's education*(avg $FemEdu_{cd}$ − avg $FemEdu_d$)* is the interaction between the education difference term and husband's own education which captures the potential assortative matching effect of education. Controls I consists of age at marriage of the wife and economic status of the wife's natal family as compared to the husband's family at the time of marriage. Controls II consists of per capita annual income of the husband's family, its assets and its rural or urban location at the time of the survey. Robust standard errors clustered at the primary sampling unit level are in paranthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Regressions weighted by survey weight of the eligible woman.

Table 2.5: Inter caste marriages and parental education

	(1) ICmarriage	(2) ICmarriage	(3) ICmarriage	(4) ICmarriage
husband's education	-0.000351	-0.000364	-0.000364	-0.0000839
	(0.000543)	(0.000545)	(0.000549)	(0.000530)
wife's education	-0.000776	-0.00117	-0.00110	-0.000886
	(0.000839)	(0.000831)	(0.000849)	(0.000820)
husband's mother's education		0.00181**	0.00186**	0.00186**
		(0.000889)	(0.000889)	(0.000874)
husband's father's education		-0.000953	-0.000932	-0.000842
		(0.000626)	(0.000635)	(0.000632)
wife's mother's education		0.00105	0.00109	0.00104
		(0.000927)	(0.000929)	(0.000917)
wife's father's education		0.000284	0.000274	0.000327
		(0.000524)	(0.000526)	(0.000514)
Controls I			√	√
Controls II				√
Caste controls	√	√	√	√
Year of marriage FE	√	√	√	√
District FE	√	√	√	√
N	22476	22251	22251	22244
R^2	0.221	0.223	0.223	0.224

Note: Linear probability results are reported. Data source is IHDS-II. Outcome is a dummy variable which takes value 1 if the marriage is inter caste, 0 otherwise. Controls I consists of age at marriage of the wife and economic status of the wife's natal family as compared to the husband's family at the time of marriage. Controls II consists of per capita annual income of the husband's family, its assets and its rural or urban location at the time of the survey. Robust standard errors clustered at the primary sampling unit level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Regressions weighted by survey weight of the eligible woman.

Table 2.6: Robustness checks: Variations in the sample of women and inclusion of interaction fixed effects

	(1) Completed education before marriage	(2) Only arranged marriages	(3) District*Year of marriage FE	(4) Logit
husband's education	-0.000197	-0.000167	-0.000675	0.000441
	(0.000538)	(0.000639)	(0.000735)	(0.0145)
wife's education	-0.000979	-0.000659	-0.000100	-0.0206
	(0.000875)	(0.000645)	(0.000829)	(0.0202)
husband's mother's education	0.00226**	0.00210**	0.00220*	0.0471**
	(0.000951)	(0.000862)	(0.00123)	(0.0184)
husband's father's education	-0.000878	-0.00110	-0.00103	-0.0208
	(0.000670)	(0.000704)	(0.000790)	(0.0168)
wife's mother's education	0.000978	0.000313	0.000633	0.0204
	(0.000968)	(0.000736)	(0.00109)	(0.0200)
wife's father's education	0.000264	0.000696	0.000550	0.0127
	(0.000539)	(0.000499)	(0.000684)	(0.0128)
Controls I	√	√	√	√
Controls II	√	√	√	√
Caste controls	√	√	√	√
Year of marriage FE	√	√	√	√
District FE	√	√	√	√
District*Year of marriage FE			√	
N	21269	16439	22244	16089
R^2	0.229	0.339	0.549	-

Note: Linear probability results are reported in columns 1 to 3. Logit results are reported in column 4. Outcome is a dummy variable which takes value 1 if the marriage is inter caste, 0 otherwise. Data source is IHDS II. Column 1 uses the sample of only those women who had completed their education before they got married. Column 2 uses the sample of only arranged marriages defined as in text. Column 3 adds interaction of district and year of marriage fixed effects to the set of district fixed effects and year of marriage fixed effects. Column 4 uses a logistic regression specification. Robust standard errors clustered at the primary sampling unit level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Regressions weighted by survey weight of the eligible woman.

Appendix

Varna, *jati*, and caste categories

According to Deshpande (2011), in the ancient Hindu society, the institution of caste was divided into initially four and later five mutually exclusive *varnas* which were hereditary, endogamous and occupation specific. They were called *Brahmins* (priests and teachers), *Kshatriyas* (warriors and the royalty), *Vaishyas* (traders, merchants and money lenders) and *Shudras* (peasants and other menial and lowly job workers). The fifth category were the *Atishudras* who did the most polluting and menial jobs. These were the formal untouchables. The *varnas* are theoretically ranked according to the following hierarchy: *Brahmins* at the top, followed by *Kshtriyas*, *Vaishyas* and then *Shudras*. The *Atishudras* were the lowliest of the low and were in fact called the *avarnas* or without a *varna*. In other words, they were excluded from the caste system.

The building blocks of the contemporary social code are *jatis*, which are subcategories of the *varnas*. However, there does not exist a one-to-one mapping of a *jati* to a *varna*. There is a lot of fluidity and ambiguity involved in their categorization due to the numerous, and in most cases, unverifiable, claims of *varna* affiliations made by the more than 3000 *jatis* in India (Deshpande 2011).

The caste categories used in this chapter are, on the other hand, administrative categories. When affirmative action policies were being formulated, *jatis* which were economically the weakest and were historically subjected to discrimination and deprivation, the so called "untouchables", were identified in a government schedule as the target group for reservation policies (Deshpande, 2011). These *jatis* are referred to as the Scheduled Castes (SC). Another government schedule identified similarly placed tribes and tribal communities for the reservation policy and they are referred to as the Scheduled Tribes or ST.

The Mandal Commission, appointed in 1979 by the then prime minister of India, Morarji Desai, recommended that the reservation policy be extended to a third group of *jatis* which were not former untouchables but were economically and educationally backward. These *jatis* are categorized as the Other Backward Classes or OBC. The residual category is often called the general category or the "Others" to mean all the castes that are not included in the Scheduled Castes (SC), Scheduled Tribes (ST) or Other Backward Classes (OBC). The IHDS is the unique data set which divides the "Others" category further into *Brahmins* and Other Forward Castes (OFC) to separate the group at the very top of the caste hierarchy.

Inter caste marriages and own education: Status exchange?

Status exchange theory argues that in a hierarchical society, out-marriage often involves an exchange of characteristics by the two parties. Typically one party exchanges its social status for some other trait of the spouse, like beauty or education. Thus, more educated blacks would marry less educated whites because they would

gain from a higher social status of their spouse (Gullickson 2006). In the Indian context, since there are more than two castes, there can potentially be many types of higher caste-lower caste unions. Hence our null results might mask interesting heterogenity across castes in the relationship between education and exogamy.

To test for the above possibility, we use the following specification:

$$\begin{aligned} ICmarriage_{id} = \gamma_0 &+ \beta_0.husband's\ caste_{id} + \beta_1.husband's\ education_{id} \\ &+ \beta_2.husband's\ education_{id}.BR_{id} + \beta_3.husband's\ education_{id}.OBC_{id} \\ &+ \beta_4.husband's\ education_{id}.SC_{id} + \alpha_0.wife's\ education_{id} \\ &+ \alpha_1.wife's\ education_{id}.BR_{id} + \alpha_2.wife's\ education_{id}.OBC_{id} \\ &+ \alpha_1.wife's\ education_{id}.SC_{id} + \sigma.X_{id} + \delta_d + \tau_t + \xi_{id}. \end{aligned} \tag{a1}$$

Here husband's caste are caste dummies for Brahmin, OBC and SC.[31] The omitted category is OFC. If status exchange takes place, it implies that compared to OFC, the marginal effect of an increase in education will be higher for OBC and SC which are ranked lower than the OFC in the caste hierarchy, while it will be lower for Brahmins who are ranked above OFC. Thus we expect β_2 to be negative while β_3 and β_4 to be positive.

The opposite will hold true for the wife because we are interacting her education with her husband's caste. Therefore, the marginal effect of an increase in the education of the wife will be positive when the husband's caste is higher than OFC (the omitted category), whereas it will be negative when the husband's caste is lower than OFC. Thus we expect α_1 to be positive while α_2 and α_3 to be negative.

The results are reported in Table A1. The first column reports coefficients when

[31] As mentioned earlier, we use only husband's caste in our specification since we do not know the caste of the wife in a couple.

year of marriage fixed effects and district fixed effects are not included. It can be seen that none of the coefficients are statistically significant here. It implies that education does not *differentially* improve the chances of an inter caste marriage for any caste as compared to the OFCs, the omitted category. In columns 2 to 4, controls are successively added to the base specification. The results do not change: none of the coefficients are statistically significant in any of the columns.

Robustness checks: Variations in religion and caste composition of samples

In this set of robustness checks, we test whether our results withstand variations in the sample which consists of only those households which have stated their religion as Hinduism, Buddhism, Jainism or Sikhism residing in the 20 major states in India, and excludes scheduled tribes. Table A2 shows the results for four such sample variations. Since caste system, as mentioned above, is theoretically a Hinduism phenomenon, in the first column in Table A2, we look at the sample of only Hindus and drop all those households who report their religions to be Buddhism, Jainism or Sikhism. In column 2, we expand the sample to include all religions in the major states because, as mentioned earlier, in our data set households belonging to all religions have reported their castes. In the next two columns, we expand the sample further to include all religions in all states and to all religions and all castes (including the STs) in all states, respectively. The results reported in all the columns are qualitatively similar to those in the main regression: the education of the spouses themselves do not matter whereas that of the husband's mother has a positive and statistically

significant association with the likelihood of an inter caste marriage[32].

Robustness check: Bound Analysis (Oster 2019)

We conduct the bound analysis to check the robustness of our coefficient on the education of the husband's mother variable to omitted variable bias. In this section we first describe the methodology briefly, and then report our results.

It is a common practice to infer about the robustness of a result to omitted variable bias by looking at coefficient movements upon the addition of controls. Oster (2019) argues that to use observables to estimate bias from unobservables, we must (a) invoke the assumption of related covariance, that is, we need to assume that the unobservables positively covary with the observables so that the observables are informative about the unobservables, *and* (b) scale the coefficient movements by movements in R^2. Building on Altonji et al. (2005) and using the assumption of related covariance, she explicitly links coefficient movements, R^2 movements and omitted variable bias. In particular, she assumes a proportional selection relationship between the observables and unobservables, and denotes this coefficient of proportionality by δ. Thus, δ essentially captures the relative strength of unobservable selection to observable selection. Using this assumption, one can calculate the bias adjusted value of the coefficient of interest, assuming a value for δ and a value for the R^2 in the hypothetical regression which controls for both observables and unobservables (R^2_{max}). If unobservables are as important as observables, then $\delta = 1$. Oster (2019) suggests that this is a reasonable upper bound for δ, that is, unobservables

[32]Apart from these samples, we ran the regressions for the following other combinations of religions, castes and states: four main religions, main states, including STs; all religions, main states, including STs; and four main religions, all states, excluding STs. Our results are robust to all these sample specifications. These results, not reported here, are available upon request.

should not be more important than the observables in explaining the dependent variable. An upper bound for R^2_{max} is equal to 1 for the case when all of the variation in the dependent variable is explained by the observables and unobservables combined. However, this may often not be the case and therefore, $R^2_{max} = min(\pi.R^2_{controlled}, 1)$ is a suggested function to arrive at an upper bound for R^2, where $R^2_{controlled}$ is the R^2 from the regression including all the observable controls and π is a multiplier[33]. This exercise will give a bias adjusted value of the coefficient of interest which can then be compared to the value of the coefficient in the controlled regression.

Assuming an appropriate bounding value for R^2, or R^2_{max}, one can also calculate the value of δ which renders the coefficient of interest zero. A value of δ greater than 1 would suggest a robust coefficient.

We carry out the bound analysis in both the ways as suggested by Oster (2019). The results are reported in Table A3. Using similar terminology, δ captures the relative importance of the unobservables with respect to the observables, and β is the coefficient of our variable of interest, that is, the years of education of the husband's mother. We first report the bias adjusted β under the assumption δ=1 and an upper bound for R^2_{max}[34]. We use the function $R^2_{max} = min(\pi.R^2_{controlled}, 1)$ and set $\pi = 1.3$ (Oster 2019). This translates to $R^2_{max} = 0.2912$ and the corresponding $\beta = 0.00159$. The bias adjusted coefficient has the same direction as that reported in our analysis up till now. We then calculate the value for δ if β were to equal zero, with the same assumption on R^2_{max}. As can be seen from the table, the value of δ comes out to be equal to 2.65. This suggests that for β to actually be statistically insignificantly

[33]Oster (2019) applies this adjustment to a host of studies, both randomized and non-randomized, as well as to constructed data to see whether the results survive. Taking randomized experiment results as the benchmark, she suggests $\pi = 1.3$ as the cutoff value at which at least 90% of the randomized results survive. We use this value of π for our tests.

[34]The term R^2max is the R^2 in the hypothetical regression which controls for both observables and unobservables.

different from zero, the unobservables must be almost three times as important as the observables. Since this seems unlikely to be the case, we deduce that our result is robust.

Finally, we also report the 95% confidence interval for the controlled β and check if the "identified set" (bounded on one side by the uncontrolled regression coefficient and on the other side by the bias adjusted coefficient of interest) lies within the confidence interval. The last two rows of Table A3 show that the identified set indeed falls within the 95% confidence interval of the controlled β. This lends further proof that the coefficient of the husband's mother's education variable is not contaminated by omitted variables bias.

Education and Female Autonomy

In this section we discuss the link between the decision making power of women and their education level using the responses to various questions under the "Gender Relations" section in the eligible woman's questionnaire in IHDS II. We run the following linear probability model:

$$Autonomy_{id} = \beta_0 + \beta_1.own\ \ education_{id} + \theta.X_{id} + \delta_d + \tau_t + \varepsilon_{id}. \tag{a2}$$

Here $Autonomy_{id}$ is a dummy variable which takes value 1 when an eligible woman i in district d has the *most say* in a particular household decision. The "Gender Relations" section asks eight such questions. The variable of interest is *own education* which is the years of education of the respondent woman. A host of control variables are included in X. Apart from the standard control variables used in this research – caste category of the household, comparative economic status of the woman's natal

family and her husband's family at the time of their marriage, per capita income and asset index of the household and its rural or urban location, X also includes the age of the respondent as well as the years of education of her husband and both parents in law. We include district fixed effects, δ_d, and year of marriage fixed effects, τ_t, in all the regressions. The results are reported in the two panels of Table A4.

In Panel A, we can see that education of the eligible woman is positively correlated to her having the most say in whether to buy an expensive item (column 2), how many children to have (column 3) and what to do if she falls sick (column 4). We see the same positive association in Panel B in decisions on whether to buy land or property (column 1), how much money to spend on a social function (column 2), what to do if her child falls sick (column 3) and, most importantly for our analysis, to whom should her children marry (column 4).

Education as a categorical variable

It is conceivable that in the marriage market, education levels of prospective brides and grooms are presented in threshold values. For example, the information may be bunched at "primary educated" or "has a bachelors degree", and not in terms of years of education. If this is the case then we may find larger coefficient sizes around these threshold values. In this section we check the robustness of our results by including education of the spouses as categorical variables. We divide the years of education into five mutually exclusive categories for both the husband and the wife. These are: Illiterate (0 years of education), Up to primary (1 to 5 years of education), Up to Secondary (6 to 10 years of education), Up to bachelors (11 to 15

years of education) and Above bachelors (16 or more years of education)[35].

The results are presented in Table A5. We see that both parts of our results hold when we introduce spouses' education as categorical variables. The education levels of the spouses themselves do not matter at any threshold. The education level of the husband's mother's education, here too, is the only one with a positive and statistically significant association with the likelihood of her son's marriage being an inter caste one.

[35]Education levels of the parents can also be included as categorical variables. However, that will mean 24 hypothesis being tested simultaneously (4 for each individual). As discussed in Section 2.6, none of the coefficients will retain their statistical significance after multiple hypothesis correction. Thus we include parents' education levels as continuous variables only.

Table A1: Own education: Status exchange?

	(1) ICmarriage	(2) ICmarriage	(3) ICmarriage	(4) ICmarriage
husband's education (β_1)	-0.000367	-0.000109	-0.0000707	0.000107
	(0.00237)	(0.00183)	(0.00182)	(0.00187)
husband's education*BR (β_2)	0.00457	0.00211	0.00210	0.00213
	(0.00388)	(0.00236)	(0.00236)	(0.00235)
husband's education*OBC (β_3)	-0.0000651	-0.000123	-0.000159	-0.0000818
	(0.00289)	(0.00237)	(0.00239)	(0.00237)
husband's education*SC (β_4)	-0.00157	-0.000915	-0.000943	-0.000884
	(0.00265)	(0.00207)	(0.00207)	(0.00207)
wife's education (α_0)	-0.00117	-0.00202	-0.00192	-0.00180
	(0.00304)	(0.00246)	(0.00250)	(0.00242)
wife's education*BR (α_1)	0.00109	0.00104	0.00109	0.000953
	(0.00378)	(0.00293)	(0.00294)	(0.00291)
wife's education*OBC (α_2)	0.0000876	0.00152	0.00149	0.00154
	(0.00323)	(0.00230)	(0.00232)	(0.00232)
wife's education*SC (α_3)	0.00286	0.00168	0.00162	0.00170
	(0.00318)	(0.00241)	(0.00241)	(0.00241)
Controls I			√	√
Controls II				√
Caste controls	√	√	√	√
Year of marriage FE		√	√	√
District FE		√	√	√
N	22476	22476	22476	22476
R^2	0.002	0.221	0.221	0.222

Note: Linear probability results are reported. Data source is IHDS-II. Outcome is a dummy variable which takes value 1 if the marriage is inter caste, 0 otherwise. The Greek letters specified in parenthesis in the first column against the variable names correspond to the notations used in equation 4. Column 1 does not include year of marriage fixed effects and district fixed effects. Controls I consists of age at marriage of the wife and economic status of the wife's natal family as compared to the husband's family at the time of marriage. Controls II consists of per capita annual income of the husband's family, its assets and its rural or urban location at the time of the survey. Robust standard errors clustered at the primary sampling unit level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Regressions weighted by survey weight of the eligible woman.

Table A2: Robustness checks: Variations in religion and caste composition of the samples

	(1) Only Hindus	(2) All religions, main states	(3) All religions, all states	(4) All religions, all castes, all states
husband's education	-0.0000800	-0.000126	0.0000413	0.0000288
	(0.000542)	(0.000470)	(0.000474)	(0.000451)
wife's education	-0.000913	-0.00106	-0.000986	-0.000760
	(0.000848)	(0.000726)	(0.000716)	(0.000705)
husband's mother's education	0.00199**	0.00191**	0.00169**	0.00146*
	(0.000928)	(0.000833)	(0.000816)	(0.000836)
husband's father's education	-0.000978	-0.000727	-0.000575	-0.000569
	(0.000637)	(0.000582)	(0.000573)	(0.000558)
wife's mother's education	0.000917	0.000457	0.000674	0.000773
	(0.000955)	(0.000839)	(0.000827)	(0.000810)
wife's father's education	0.000356	0.000501	0.000388	0.000278
	(0.000527)	(0.000476)	(0.000473)	(0.000469)
Controls I	√	√	√	√
Controls II	√	√	√	√
Caste controls	√	√	√	√
Year of marriage FE	√	√	√	√
District FE	√	√	√	√
N	21309	25693	26707	29030
R^2	0.226	0.198	0.230	0.220

Note: Linear probability results are reported. Outcome is a dummy variable which takes value 1 if the marriage is inter caste, 0 otherwise. Data source is IHDS II. Column 1 uses the sample of only Hindus in the main states. Column 2 uses the sample of all religions, excluding STs, in the main states. Column 3 uses the sample of all religions, excluding STs in all states. Column 4 includes all religions, all castes including STs in all states. Robust standard errors clustered at the primary sampling unit level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Regressions weighted by survey weight of the eligible woman.

Table A3: Bound analysis

Uncontrolled β	0.00221
(R^2)	(0.001)
Controlled β	0.00186
(R^2)	(0.224)
β **for** $\delta = 1$, $R^2_{max} = 0.2912$	0.00159
δ **for** $\beta = 0$, $R^2_{max} = 0.2912$	2.65
Identified set	[0.00159, 0.00186]
95% Confidence interval	[.0001462, 0.003575]

Note: Bound analysis results are reported. Here $R^2_{max} = min(\pi.R^2_{controlled}, 1)$, $\pi = 1.3$. The Uncontrolled regression controls only for the education of the husband's mother, the controlled regression includes the full set of education variables and control variables. Data source is IHDS II.

Table A4: Education and female autonomy

Panel A				
	(1) What to cook on a daily basis	(2) To buy an expensive item	(3) How many children should she have	(4) What to do if she falls sick
own education	-0.00121	0.00207***	0.00489***	0.00172*
	(0.00134)	(0.000667)	(0.00118)	(0.000912)
Panel B				
	(1) To buy land or property	(2) How much money to spend on a social function	(3) What to do if her child falls sick	(4) To whom her children should marry
own education	0.00181***	0.00342***	0.00505***	0.00168**
	(0.000590)	(0.000864)	(0.00109)	(0.000760)
Controls I	√	√	√	√
Controls II	√	√	√	√
Caste controls	√	√	√	√
Year of marriage FE	√	√	√	√
District FE	√	√	√	√

Note: Linear probability results are reported. Data source is IHDS-II. Outcome is a dummy variable which takes value 1 if the eligible woman has the most say in that household decision. Controls I consists of age of the respondent and years of education of her husband, her mother in law and her father in law. Controls II consists of the economic status of the wife's natal family as compared to the husband's family at the time of marriage, per capita annual income of the husband's family, its assets and its rural or urban location at the time of the survey. Robust standard errors clustered at the primary sampling unit level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Regressions weighted by survey weight of the individual.

Table A5: Education as categorical variables

	(1) ICmarriage	(2) ICmarriage	(3) ICmarriage
wife up to primary	-0.00612	-0.00628	-0.00517
	(0.00677)	(0.00683)	(0.00666)
wife up to secondary	-0.0107	-0.0104	-0.00838
	(0.00965)	(0.00965)	(0.00916)
wife up to bachelors	-0.0124	-0.0113	-0.00910
	(0.0104)	(0.0105)	(0.0102)
wife above bachelors	0.00565	0.00816	0.00877
	(0.0188)	(0.0190)	(0.0189)
husband up to primary	-0.00261	-0.00284	-0.00202
	(0.00834)	(0.00827)	(0.00808)
husband up to secondary	-0.00620	-0.00640	-0.00432
	(0.00651)	(0.00656)	(0.00622)
husband up to bachelors	-0.0102	-0.0101	-0.00699
	(0.00660)	(0.00661)	(0.00654)
husband above bachelors	-0.0190*	-0.0185*	-0.0155
	(0.0107)	(0.0107)	(0.0107)
husband's mother's education	0.00165*	0.00169*	0.00172*
	(0.000949)	(0.000948)	(0.000927)
husband's father's education	-0.000882	-0.000860	-0.000774
	(0.000651)	(0.000660)	(0.000652)
wife's mother's education	0.000866	0.000895	0.000875
	(0.000851)	(0.000854)	(0.000851)
wife's father's education	0.000309	0.000307	0.000355
	(0.000532)	(0.000533)	(0.000518)
Controls I		√	√
Controls II			√
Caste controls	√	√	√
Year of marriage FE	√	√	√
District FE	√	√	√
N	22265	22265	22258
R^2	0.223	0.224	0.224

Note: Linear probability results are reported. Data source is IHDS-II. Outcome is a dummy variable which takes value 1 if the marriage is inter caste, 0 otherwise. The omitted group for both the spouses is illiterate category. Controls I consists of age at marriage of the wife and economic status of the wife's natal family as compared to the husband's family at the time of marriage. Controls II consists of per capita annual income of the husband's family, its assets and its rural or urban location at the time of the survey. Robust standard errors clustered at the primary sampling unit level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Regressions weighted by survey weight of the individual.

Chapter 3

What can(not) explain the gap? Evidence and decomposition of gendered stream choice in India

3.1 Introduction

Gender gap in earnings is well established in the economics literature, both in the context of developed (Rubery 1992; O'Neill 2003; Plantenga et al. 2006; Blau and Kahn 2017; Boll and Lagemann 2018) as well as developing countries (Ashraf and Ashraf 1993; Brown et al. 1999; Maurer-Fazio et al. 1999; Rendall 2013; Chi and Li 2014; Guimarães and Silva 2016.), including India (Reilly et al. 2005; Menon and Van der Meulen Rodgers 2009; Khanna 2012; Das 2012; Duraisamy and Duraisamy 2016; Deshpande et al. 2018.). Even in 2014, annual earnings of full-time female workers in the USA were only about 79% that of male workers (Blau and Kahn

2017), while the cross-country unadjusted gender gap in average wages in Europe was 14.2% (Boll and Lagemann 2018). The gap stood much higher at 49% in India in the year 2009-10 (Deshpande et al. 2018). While a number of explanations have been extended to explain this earnings gap, occupational segregation has emerged as a major explanatory factor (See, for example, Dolado et al. (2002); Hegewisch et al. (2010); Hegewisch and Hartmann (2014); Blau and Kahn (2017).). In particular, male dominated occupations related to Science, Technology, Engineering and Mathematics (STEM) fields have substantial earnings premium while women are over represented in lower paying jobs like nursing and teaching (See James et al. (1989); Grogger and Eide (1995); Dolton and Vignoles (2002); McGuinness (2003); Buonanno and Pozzoli (2009); Webber (2016); Belfield et al. (2018); Jain et al. (2018); Dahl et al. (2020).). Why are so few women employed in STEM related occupations despite a clear economic advantage in these fields? It could be either because fewer women graduate in STEM or because women graduate in STEM but drop out of STEM occupations, or a combination of both. In most countries, however, specialization into STEM, or more broadly, into Science and non-Science streams happens even earlier, at the school level itself. In this chapter, we look at the very first stream choices made by students at the school level in the context of India using three cohorts of students under the single-largest education board[1] with an all-India presence – the Central Board of Secondary Education (henceforth CBSE).

A rich literature is devoted to exploring the college major choice of students (Turner and Bowen 1999; Montmarquette et al. 2002; England and Li 2006; Dickson 2010; Riegle-Crumb and King 2010; Riegle-Crumb et al. 2012). Recent work has shifted focus to look at stream choices made earlier, at the school level (Baram-

[1]An education board in India is defined by its jurisdiction. There are three national boards and several state boards in India (Cheney et al. 2005).

Tsabari and Yarden 2011; Ajayi and Buessing 2015; Friedman-Sokuler and Justman 2016; Justman and Méndez 2018; Rapoport and Thibout 2018; Landaud et al. 2020). In the Indian context, however, the literature is patchy and limited in scope. While Mattoo (2013) and Gautam (2015) have limited geographic scope, Chanana (2000, 2007) only look at aggregate statistics. Chakrabarti (2009) and more recently, Prakasam et al. (2019) use nationally representative data sets to look at major choices at the college level. In a recent paper, Sahoo and Klasen (2020) look at stream choices at the school level using the nationally representative Indian Human Development Survey. To the best of our knowledge, this is the first attempt to use multiple cohorts of a large scale dataset to establish and subsequently decompose the gender gap in stream choice in India. Using regression and linear decomposition techniques, we decompose this gap to estimate how much of it can be "explained" by an expansive set of explanatory factors that have been proposed in the economics, sociology and psychology literature. In particular, we evaluate how much of the gender gap is accounted for by a difference in student ability, their cohort peers, their immediate seniors and their socioeconomic status. We conclusively show that a difference in student ability, the earliest explanation offered, accounts for less than 10% of the gender gap we observe in various subjects. A difference in cohort peer attributes also does not explain any statistically significant portion of the gap. Instead, we propose a novel way to use the immediate seniors of students to elicit measures of role model and chilly climate, and find that these are the largest explanators of the existing gender gap in stream choice.

The National Policy on Education, 1968, recommended a common $10 + 2 + 3$ educational structure for the entire country, with 10 years of schooling up to the secondary level (matriculation), followed by 2 years of higher secondary schooling

and three years of graduation. While the duration of graduation varied, the $10 + 2$ system has been followed in all schools since then. Under this structure, students are required to choose four specialized subjects of study after matriculation (or class X) for the next two years of higher secondary education (classes XI and XII)[2]. Since we are looking at choices made at the school level, the classification of subjects into STEM and non-STEM can be misleading for two reasons. First, the subjects offered after matriculation are more basic like Physics and Chemistry rather than Engineering. Second, "Science" in STEM incorporates both Technical sciences (like Engineering) and Life sciences (like Medicine and Anthropology), but the distinction is not sharp at the school level, again because the subjects offered are at the basic level. For example, almost 99% of students in our dataset who opt for Biology opt for Physics and Chemistry as well. Therefore, we find it more prudent to look at individual subjects and subject combinations most commonly offered by the schools. In particular, we look at three subjects/subject combinations in this chapter: Mathematics, Biology and Physics-Chemistry-Mathematics (PCM) combination. We find that gender differences are the starkest in these subjects. In addition, these subject choices lead to the most economically rewarding major choices in college, as noted in the beginning[3].

We begin by quantifying the gender gap in various subjects offered in schools after matriculation and find a clear gender divide in our dataset along the same lines as observed in the literature. For example, on an average, boys are 19.13 percentage points more likely than girls to take-up Mathematics in class XI and 20.61 percentage

[2]The most frequently offered subjects by CBSE schools in class XI in our dataset are: Mathematics, Physics, Chemistry, Biology, Computer Science, History, Political Science, Geography, Economics, Hindi, English, Business Studies and Accounts. CBSE also offers other regional languages, Music and Fine Arts and a range of vocational subjects.

[3]It must be noted that Mathematics can be a subject choice even without PCM; for example, it can be chosen with the Commerce stream (combination of Business Studies and Accounts) as well.

points more likely to take-up PCM. Girls, on the other hand, are 11.18 percentage points more likely than boys to take-up Biology. In general, a higher proportion of girls takes up Biology, Economics, Political Science and History, while boys are more likely to take-up PCM and Computer Science.

Next we proceed to dissect this gap. We go beyond student ability and explore the literature for other plausible reasons extended for the gender gap in stream choice. Our dataset allows us to study a large number of these explanatory factors. We group them into four broad heads: Ability, Cohort peers, Immediate seniors and Socioeconomic characteristics. For each of these heads, we first show descriptive statistics on how it is distributed between boys and girls, and how they relate to stream choice. Then we examine them under a regression framework using Linear Probability Models and a decomposition framework using Oaxaca Blinder decomposition technique.

Based on some early life studies which find gender gap in Mathematics scores favoring boys (Penner and Paret 2008; Fryer Jr and Levitt 2009; Wai et al. 2010), this difference is often cited as the explanation for the observed gender divide in stream choice. This claim has been falsified in multiple studies in the context of developed countries (Dickson 2010; Riegle-Crumb and King 2010; Riegle-Crumb et al. 2012; Rapoport and Thibout 2018; Friedman-Sokuler and Justman 2016; Justman and Méndez 2018). We test this in the context of a developing country using a student's class X score as a proxy for her ability. A more nuanced explanation explored in the literature is comparative advantage in the relevant subject. It is argued, for example, that boys have a comparative advantage in Mathematics over languages while girls have a comparative advantage in life sciences over other technical sciences, and thus we observe the existing gender divide in stream choices (Park et al. 2007; Valla and Ceci 2014). For this, we use class X scores in the relevant subjects to

construct comparative advantage variables. These two variables are considered under the ability head of our explanatory variables.

Under the cohort peers head, we look at the properties of peers of a student at her class X school-cohort level. We explore the argument that males and females may have very different responses to a given set of peers (Gneezy et al. 2003; Gneezy and Rustichini 2004; Niederle and Vesterlund 2007; Gneezy et al. 2009; Fletschner et al. 2010). The reasons for this are said to be rooted in differential confidence[4] and risk taking behaviour [5] across genders. To see this, we include two measures of the properties of cohort peers. First we look at the gender composition of the cohort peers of student to gauge whether girls and boys behave differently under a given share of own gender students in the cohort. Second we look at peer performance variables or the average class X performance of cohort peers of a student to check whether girls and boys perform differently among peers of a given quality.

Next, we employ a novel way to utilize the immediate seniors that students had in class X to elicit some other explanatory factors propounded in the literature. Role models (or lack thereof) have been widely argued to shape how students perceive the viability of a prospective field. Generally, only teachers and instructors are considered potential role models for pupils (Bettinger and Long 2005; Hoffmann and Oreopoulos 2009; Paredes 2014; Fairlie et al. 2014; Bottia et al. 2015). However, a student interacts with many others in an institutional environment. Just as cohort peers may influence the behaviour of a student, her seniors may also be potential role models. We here examine this potential channel by examining how the subject choices of the immediate seniors of students correlate with their own subject choices. Another possible explanation proposed in the sociological and psychological literature

[4]See Jakobsson (2012); Pirinsky (2013); Sarsons and Xu (2015).
[5]See Charness and Gneezy (2012); Hardies et al. (2013).

is the phenomenon of "Chilly Climate"(Clark Blickenstaff 2005). It says that females are shy of choosing male dominated fields because they face a hostile environment there. If too few fellow females are present in a Mathematics class, then there are higher chances of covert and overt discrimination, or a general feeling of being at a loss (See, for example, Sadker and Sadker (1986); Fouad et al. (2011); Lordan and Pischke (2016); Tellhed et al. (2017); Wu (2017).). The same could be applicable to males in a Biology class dominated by females. We propose that students may form an idea of a prospective Mathematics class, for example, by looking at the gender composition of the Mathematics class of their seniors. We are able to study these two explanatory factors by examining the behaviour of students' seniors under this head. To the best of our knowledge, we are the first to use information on a student's seniors to study the role model and chilly climate aspect in stream choice.

Finally, we probe how much of the gender gap can be attributed to socioeconomic characteristics of students. For this, we use their caste status[6], their annual family income and their single child status as variables signaling their socioeconomic status.

It becomes clear early on in our analysis that the distribution of ability and peer related attributes does not differ dramatically across girls and boys. This is first evident from our descriptive statistics. For instance, we find that girls score better than boys in each subject in class X, except for Mathematics, where they score lower by 0.04 standard deviations. They have a higher female share in their cohort on average than boys, but lower performing cohort peers. However, there are large differences between the genders in the attributes related to seniors. Girls have far fewer own gender seniors who opted for Mathematics and PCM than boys, while boys have much lower values of own gender seniors who opted for Biology. Similarly,

[6]Refer to the Appendix to Chapter 2 for details on the caste system in India.

the senior Mathematics and PCM classes of girls have half the share of own gender students as compared to the senior Mathematics and PCM classes for boys. The opposite is observed for girls and boys in Biology. The returns to these attributes are estimated using regression analysis. For each subject, we run a linear probability model with a binary dependent variable which takes value one if a student opted for that subject in class XI. The right hand side variables include a dummy variable for being a female student. Measures of comparative advantage in ability have larger and more statistically significant coefficients than the absolute ability terms. The attributes related to the cohort peers of students have coefficients that are small in size and they do not affect the size of the female dummy coefficient. The senior related variables, however, reduce the size of this coefficient by the largest amount. This gives a prelude to what we find in the formal decomposition exercise.

We report three broad findings. One, role model and chilly climate aspects of a student's immediate seniors in school are the largest explanators of the gender gap in Mathematics, PCM and Biology. If girls had the share of own gender students in the senior Mathematics and PCM classes like boys, and the share of own gender seniors choosing Biology like boys, the gender gap in these subjects would have closed by 24%, 16% and 18%, respectively. Two, for Mathematics and PCM, a comparative advantage in Mathematics vs English is the second largest contributor to explain the gender gap. For Biology, a comparative advantage in Science vs. Mathematics emerges as the second largest contributor. Three, peer composition and peer performance variables do not explain any statistically significant portion of the gender gap in any of the three subjects.

Our contribution to the literature is fourfold. First, we use a newly available administrative results dataset of the census of students under the largest national

level education board in India. The rich student level data allows us to study fine differences within a broad stream. For example, we are able to look at mathematical and non-mathematical subjects within the Science stream. This is particularly important in the Indian context where most secondary large datasets have information only at the broad stream level like Science, Arts and Commerce. Similarly, where most datasets only have information on the overall grades (first or second division) and results (pass/fail) of students, we know the exact subject-wise scores of each student. This gives an upper edge in exploring individual level outcomes using individual level controls.

Second, while a substantial work has been done on stream choice in developed countries, most papers focus on only a particular explanatory factor, like ability or classroom peers. This chapter is the first one to present multiple factors covering a broad spectrum of explanations in a unified framework empirically. We utilize an expansive set of observable factors around a student to account for the gender gap present in stream choice.

Third, we implement a novel way to elicit role model and chilly climate aspects of stream choice using school seniors of students. Previous work has only used teachers and instructors to construct measures of role models for students, and the composition of classroom peers to measure "chilliness" of the chosen course of study. To the best of our knowledge, this is the first attempt to use a student's seniors, whom the students observe *before* they make their own decisions regarding stream choices, to construct these measures. Most importantly, we also find these explanatory factors to be the largest contributors to explain the gender gap in the take-up of Mathematics, PCM and Biology.

Fourth, this is also the first work to rigorously establish and subsequently decom-

pose the gender gap in stream choice at the school level using multiple cohorts of student level data in the context of India. The existing work in the Indian context is patchy and limited in geographic scope. We are able to fill this gap using rich detailed data of over 2 million students.

The rest of the chapter is organized as follows. In Section 3.2 we describe the institutional background followed by a description of the dataset in Section 3.3. In Section 3.4 we provide a descriptive analysis with respect to each explanatory factor and each of the three subjects. Section 3.5 briefly discusses the methodologies used in the chapter. Section 3.6 shows the results of the regression and the decomposition exercises. Section 3.7 offers a brief discussion of the results and its implications, and Section 3.8 concludes.

3.2 Institutional Background

The education system in India follows a 10+2+3 structure recommended by the National Policy on Education, 1968. It comprises of 10 years of schooling up to the secondary level culminating into matriculation in class X. This is followed by 2 years of higher secondary schooling in class XI and XII, and then 3 years of graduation. In general, the duration of graduation varies depending on the type of course and degree. The duration of schooling, however, uniformly follows the 10+2 pattern across the country (Cheney et al. 2005).

A school in India is affiliated to a board of education. A board is defined by its jurisdiction and follows a common curriculum across all affiliated schools[7]. There are three national and several state boards in India (Cheney et al. 2005). The Central

[7]Though a school's syllabus is the responsibility of the board it is affiliated to, in theory, it must be aligned with the National Curriculum Framework, 2005 (Anderson and Lightfoot 2019).

Board of Secondary Education (CBSE) is the single-largest education board in the country with an all-India presence. As of 2019, there were 21271 schools affiliated to CBSE in India and 228 schools in 25 foreign countries (`cbse.nic.in`).

All schools under the CBSE follow the same curriculum till class X, except for the choice of languages (http://cbseacademic.nic.in/curriculum_2021.html)). All education boards conduct board level standardized examinations at the end of secondary school in class X, and then at the end of higher secondary school in class XII. These examinations, commonly called board examinations, have common question papers and evaluation guidelines across the board. Both the board examinations, especially the class XII board examinations, are high stake examinations because their results are used for admission to various colleges and institutions for higher education.

Under the CBSE, after studying a common syllabus till class X (two languages, Mathematics, Science[8] and Social Science (http://cbseacademic.nic.in/curriculum_2021.html)), students have to choose one language and four specialized subjects for the next two years of study (classes XI and XII). CBSE offers a wide range of scholastic and co-scholastic subjects to choose from[9]. The choices offered by a school, however, may be limited due to resource constraints on part of the school. The most common subjects opted by students in our dataset are Mathematics, Physics, Chemistry, Biology, Computer Science, History, Political Science, Geography, Economics, Hindi, English, Business Studies and Accounts. Though a student can choose any set of subjects from the available options, there are some combinations chosen most frequently. These include Physics-Chemistry-Mathematics (PCM), Physics-Chemistry-Biology (PCB),

[8]It may be noted that till class X, students are taught the subject "Science". This is then split up into Physics, Chemistry and Biology in class XI. Throughout the chapter we use Science to mean the subject taught in class X.

[9]For the complete list of subjects offered by CBSE, please refer to `http://cbseacademic.nic.in/curriculum_2021.html`.

History-Political science (Arts) and Business Studies-Accounts (Commerce). With this institutional setting in the background, the next section describes our dataset is detail.

3.3 Data

We use three cohorts of newly available results data from the Central Board of Secondary Education in India[10]. As mentioned above, CBSE is the largest national school board in the country with an all-India presence with 21271 schools in India and another 228 schools in 25 countries affiliated to it. The board conducts two national level standardized examinations every year, one for class X and the other for class XII students. Over 1.2 million students appear for class X board examinations annually, while over 1 million students appear for class XII board examinations on an average. The stream of study chosen after matriculation is a major determinant of the field of study a student can choose in college or university taking admission via various national and state level competitive examinations. We study the stream choices of three cohorts of students who took the standardized board examinations under the CBSE at both class X and XII level.

For each student, we have board examination results data for each subject in class X (2 languages, Mathematics, Science and Social Science) and for each subject in class XII which they opted for after class X. We have their exact scores out of 100 in each of these subjects as well as their grades ranging from A to E, along with the overall score, grade and final result (pass/fail). Thus, for each class XII student, we know the subjects they opted for their higher secondary schooling and

[10]This data was obtained as part of a collaboration with the CBSE officials and is not publicly available.

their scores in those subjects. We match these class XII students to their class X results using a unique roll number assigned to each class X student in each academic year. From a total of 3,134,622 class XII students in three cohorts from years 2014, 2015 and 2016, 2,403,955 are matched with their class X results from years 2012, 2013 and 2014, respectively. The unmatched 730,665 or 23.31% students, that is, the students who did not have a CBSE class X roll number, are the students who have migrated to CBSE in class XI from other school boards. We restrict our analysis to these 2,403,955 students who appeared for both class X and XII board examinations under the CBSE.

In addition to subject scores and grades, we also have information on the gender, date of birth, caste status, annual family income, single child status of students and the type of board exam chosen in class X[11]. The data also tells us the type of their school administration (public, private, and so on). We also have a school identifier for each student at both class X and XII. Note that the schools attended by students in class X and XII may not necessarily be the same. This is mostly because many schools only have classes till class X. In fact 30.02% of students changed schools after matriculation (class X) in our dataset.

Table 3.1 reports the summary statistics for the variables available in the CBSE data. Out of a total of 2,403,955 students, around 44% are girls. The mean age of students in class X is 16.68 years. Almost 73% of the students belong to the unreserved General caste category, while 7.31% belong to Scheduled castes (SC), 3.33% to Scheduled tribes (ST) and the remaining 16% belong to Other Backward Classes (OBC).

In class XII, the largest majority of students (68%) are enrolled in private schools[12].

[11]The CBSE gives an option to choose between an externally conducted examination or a school based one for class X students.

[12]Schools in India can be government owned (Government), privately owned but financially aided

The Government schools come second enrolling close to 20% of students in our sample. More than 62% of class X students opt for school based board examination. The mean annual family income of students is 292,328.8 INR with a very large variance. Around 6% students are single children in their families and 0.20% have a form of disability. The mean score of students in class X is 348.72 out of 500 and in class XII is 329.91 out of 500.

3.4 Descriptive Analysis

In this section, we discuss in detail each of the explanatory factors we consider for the decomposition exercise. We describe the variables used to measure the factors, their differential distribution across girls and boys and their correlation with the subject choices of students.

3.4.1 Overall stream choice

Though we take an in-depth look into the choice of only three subjects in this chapter – Mathematics, Biology and the Physics-Chemistry-Mathematics (PCM) combination, this subsection gives a general view of the gender divide with respect to all subjects in class XI.

Table 3.2 reports the pattern of take-up of the most common subjects offered to CBSE students after matriculation. Column 1 specifies the subject/subject combination, column 2 shows the percentage of all students who opted for that subject in class XI. column 3 shows what proportion of those who opted for a subject were

by the government (Private Aided), or privately owned by individuals, trustees or societies (Independent)(Anderson and Lightfoot 2019). The Jawahar Navodaya Vidyalayas and Kendriya Vidyalayas are also government owned schools.

girls. Finally, columns 4 and 5 show the shares of girls and boys, respectively, who opted for the given subject.

The gendered pattern in the choice of subjects comes out clearly from Table 3.2. While more than 53% of boys take-up Mathematics after class X, only around 35% of girls do so. The gap is slightly muted but still close to 13 percentage points in case of Physics and Chemistry, and around 5.5% in Computer Science. In contrast, the gender gap favours the girls in case of Biology, History, Political Science and Hindi. It is the largest, at around 12 percentage points, for Biology and around 7 to 8 percentage points for History, Political Science and Hindi. For the rest of the subjects, namely Geography, Economics, English, Business Studies and Accounts, the difference in the take-up rate between girls and boys is 5 percentage points or below.

The last three rows of Table 3.2 show subject combinations[13]. It can be seen that there is an over 20 percentage points gap in the take-up of PCM in favour of boys. The difference is much smaller, and in favour of girls, in case of Arts and Commerce: 5 percentage points and 2 percentage points, respectively.

Figure 3.1 presents another way of quantifying the gender divide in stream choice. Following Turner and Bowen (1999), it plots values of the Dissimilarity Index (DI) for subject choice in each of the three cohorts in our data. The Dissimilarity Index captures the proportion of students who would need to change subjects in order for the distribution in each subject to look like the gender distribution in the entire set of students under study. For example, there are 44% girls in our sample. A DI value of 1 indicates complete segregation. A value of 0 indicates that there are

[13]It must be reiterated that these subject combinations have been created by the authors to match the most commonly taken subject combinations in CBSE. The board does not require students to opt for subjects from a combination package and they can choose any five subjects from the available subjects.

44% girls and 46% boys in each of the subjects, just like the proportions in the overall sample. Figure 3.1 shows that the DI hovers around a high value of 0.5 for all subjects and there is little movement across three years. The implication is that as much as half of the students will need to change subjects for there to be gender parity across all subjects. The DI is much lower if we look at only subject combinations, around 0.15, implying that around 15% of students will need to rearrange into the three combinations of PCM, Arts and Commerce to achieve gender parity in these combinations. This is because the subject combinations include individual subjects in which the gender gap is much smaller.

It is clear from Table 3.2 that the largest differences in take-up between girls and boys exists in Mathematics and PCM in favour of boys, and in Biology in favour of girls. From here onward in this chapter, we focus on these three streams only. Figure 3.2 plots the rate of take-up for the three cohorts for Mathematics, PCM and Biology. It can be seen that there is little movement in these plots over three years. There is an almost constant gap in all three subjects across the three years which is very much in line with what was observed in Table 3.2. We now move on to discuss each of the four categories of explanatory factors in detail.

3.4.2 Ability

It was briefly discussed in the introduction that multiple studies have established that a difference in ability across girls and boys now accounts for a negligible portion of the gender gap in stream choices (Turner and Bowen 1999; Dickson 2010; Riegle-Crumb and King 2010; Riegle-Crumb et al. 2012; Friedman-Sokuler and Justman 2016; Rapoport and Thibout 2018; Justman and Méndez 2018). As a matter of fact, as we progress towards more recent studies, lesser and lesser portion of the gender gap

is explained by a difference in ability. For example, Turner and Bowen (1999) find, using a Blinder Oaxaca decomposition, that Scholastic Aptitude Test (SAT) scores explain as much as 45% of the gender gap in the take-up of Mathematics-Physical Sciences as college major in the USA. In contrast, using the same decomposition technique, Justman and Méndez (2018) show that Australia's National Assessment Program-Literacy and Numeracy (NAPLAN), along with other control variables does not explain the gap in Physics, Information Technology and Specialist Mathematics at all.

Table 3.3 shows score comparison between girls and boys in our dataset. We use each student's class X score in the CBSE board examination as a proxy for their ability. Since students make subject choices immediately after class X, these scores make justifiable candidates. In addition, they are standard across all schools under CBSE. Thus they can be used even if students changed schools post matriculation. There are five subjects in class X: Mathematics, Science, Social Science, and two languages. While 99.83% of students have English as one language in class X, 83.76% of students have Hindi. Table 3.3 reports the scores in these five subjects for girls and boys. Column 2 gives the overall means, column 3 gives the means for girls and column 4 for boys. The last column gives the difference between the means for girls and boys in terms of the standard deviation of the overall sample (value in column 2).

It is evident from Table 3.3 that the differences between the average scores of girls and boys are very small: only a small fraction of the corresponding standard deviations. In fact, girls outperform boys in every subject, except in Mathematics, where they lag behind by 0.04 of a standard deviation. Figure 3.3 goes beyond the mean and compares boys and girls across the entire score distribution. It plots the

ratio of boys to girls in each decile of their class X score distribution. Each point in the graph is the ratio of the proportion of boys in a decile to the proportion of girls in that decile. A value above 1 implies that boys are over represented in that decile. It can be seen from the figure that for class X total score deciles, boys are over represented in the deciles below the 7th, while girls are more in the upper 3 deciles. However, for Mathematics scores, boys are over represented in the upper 2 deciles. The pattern for Science is less clear, but girls are over represented in the last decile.

Having formed an idea of the trends and differences in scores across genders, the next set of figures explores how they relate to students' stream choices. For each graph in Figure 3.4, we calculate score deciles for boys and girls separately. Then we calculate the proportion of boys (girls) in each deciles who opted for the given subject in class XI. Finally, we plot these proportions against the class X score deciles.

Beginning with the top panel, Figures 3.4a, 3.4b and 3.4c show the rate of take-up of Mathematics, PCM and Biology over class X *total* score deciles. Firstly, the take-up increases with score for both girls and boys for all three streams, implying that ability is a positive predictor of these subjects in general. Secondly, the gap in stream choice exists at virtually all deciles of the total score distribution, strongly implying that total scores will be poor explanators of the gender gap in stream choice. For Mathematics and PCM, the gap starts low at the lower deciles, then stabilizes around the 4^{th} decile. For Biology, the gap is flipped in favour of girls and is practically zero till the 2^{nd} decile. Beyond that, however, the gap increases with increasing score, mainly because the curve for boys flattens after around the 6^{th} decile.

In the next two panels, we replace the class X total score deciles with score deciles

of the relevant subject in class X to see if these do a better job of explaining the gender gap. Figures 3.4d and 3.4e in the second panel plot the gender wise take-up rates of Mathematics and PCM over class X *Mathematics* score deciles. We see a very similar pattern as seen in the previous panel: the gap starts low, stabilizes around the 4^{th} decile, and exists throughout the Mathematics score distribution. The last panel shows the take-up rates for PCM and Biology over class X *Science* score deciles. Here too, the graphs resemble the ones in the top panel. Figure 3.4g for Biology, for example, shows that the gap is negligible in the beginning and widens with increase in score.

The discussion so far in this section suggests that while ability variables are positively correlated with take-up of Mathematics, PCM and Biology, they are unlikely to account for a major portion of the gender gap. This is because the score distributions do not differ substantially between genders and moreover, the gender gap exists at every point in the score distribution. A more nuanced variant of the ability related explanations is the concept of Comparative Advantage in ability which we take up next.

Comparative advantage in ability borrows from the concept of comparative advantage in international trade. It means that boys do better in Mathematics and mathematically oriented subjects compared to subjects involving language and verbal skills. Girls, on the other hand, have an advantage in language and life sciences. It is this comparative advantage in ability that drives the gendered pattern of stream choices observed across countries (Park et al. 2007; Valla and Ceci 2014). Many recent studies have looked at the role such differential ability advantage plays in students' stream choices and found that the differences are more cultural than biological. For example, Friedman-Sokuler and Justman (2016) find no evidence that boys' com-

parative advantage in Mathematics drives their higher take-up of the subject in a sample of Israeli schools. Justman and Méndez (2018) show that comparative advantage exists but not at the gender level, rather between students with and without an English background in Victoria, Australia. In fact, Pope and Sydnor (2010) show that comparative advantage varies significantly across the US, suggesting again that culture and environment greatly affect the observed differences in abilities.

The source of the difference notwithstanding, existence of comparative advantage in a subject over others may still drive stream choice across genders. We first quantify the comparative advantage in ability in our data. Then we describe how it correlates with the take-up of our subjects. Table 3.4 accomplishes the first part. Since we are comparing across subjects, we convert absolute scores into standardized scores[14]. The first set of rows in Table 3.4 gives the mean standardized scores of class X scores for girls and boys. As also seen in Table 3.3, girls score above the mean in all subjects except in Mathematics. For the mean gap in scores in the next set of rows, we simply subtract the standardized score in English, for example, from that in Mathematics and average it over girls and boys separately. We see that boys indeed have a statistically significantly higher comparative advantage in Mathematics and Science over language compared to girls: boys score better in Mathematics and Science compared to English on an average (difference is positive). The opposite is true for girls. It is interesting to note that boys also have a comparative advantage in Mathematics compared to Science. Girls do better in languages compared to Mathematics and Science, but they also do better in Science compared to Mathematics. This could potentially account for some of the gap in the take-up of Biology, the

[14]Scores are standardized by subtracting the cohort mean from a student's own score and dividing the difference by the cohort standard deviation. The resultant standardized scores will have a mean of 0 and standard deviation of 1 within a cohort.

non-mathematical Science course. Finally, the last set of rows show that the above pattern exists in the overall sample as well: a higher proportion of boys secured higher scores in Mathematics and Science than in English, while a higher proportion of girls secured higher scores in Science than in Mathematics.

Figure 3.5 shows how this comparative advantage in particular subjects is associated with the subsequent choice of those subjects. Figure 3.5a has Mathematics as the subject choice and 3.5b has PCM with deciles of comparative advantage in Mathematics vs. English on the x -axes. Figure 3.5c has the take-up of Biology on the y-axis and deciles of comparative advantage in Science vs. Mathematics on the x-axis. For all three subjects, the take-up, on average, has a positive relationship with a comparative advantage in the relevant subject. However, the gender gap in stream choice remains intact for all deciles. It is quite stable for across deciles for Mathematics, while it slightly widens after the 8^{th} deciles in case of PCM. For Biology, the gap maximum at the 6^{th} decile, and reduces slightly beyond that.

To summarise, while ability correlates positively with a higher take-up of Mathematics, Biology and PCM, the gender gap in the take-up is undiminished across the entire distribution of scores. The same is true for comparative advantage in ability. A comparative advantage in the relevant subject is associated with a higher take-up of the subject post matriculation but the gender gap persists even among students with the comparative advantage.

3.4.3 Cohort Peers

A number of studies have shown that behavioural differences exist between males and females. Females have been shown to exhibit less confidence, lower competitiveness and higher risk aversion (Gneezy et al. 2003; Gneezy and Rustichini 2004;

Niederle and Vesterlund 2007; Gneezy et al. 2009; Fletschner et al. 2010; Charness and Gneezy 2012; Jakobsson 2012; Hardies et al. 2013; Pirinsky 2013; Sarsons and Xu 2015). An implication of these differences is that men and women may behave differently among the same peers. Undoubtedly, the education literature has built upon this observation and explored whether peer composition affects female performance differentially (Marsh et al. 2008; Hunt 2016; Anelli and Peri 2019; Fischer 2017; Kugler et al. 2017; Bostwick and Weinberg 2018; Astorne-Figari and Speer 2019; Landaud et al. 2020). Hunt (2016) and Bostwick and Weinberg (2018) show, for example, that a lower proportion of females among peers can negatively impact a woman's outcomes. Fischer (2017) and Landaud et al. (2020), on the other hand, show that higher performing peers can be detrimental to female performance and choices. We explore these two aspects of peers in our context. Peers in our analysis are defined at the class X school-cohort level.

First, Table 3.5 describes the gender composition in our data. We look at gender composition at class X level since stream choice decisions are potentially based on factors present at that time. Panel A reports that out of a total of 12,685 schools in our data, 1,046 are all-girls schools, 1,345 are all-boys schools and a majority 10,474 are co-educational schools. The average female share in cohort is 44% in the overall sample and close to 42% in the sample of co-educational schools. Panel B shows the gender composition of peers in the cohort of an average girl and boy in the data. Overall, girls have a higher proportion of girls in their peer group at school, and boys have higher proportion of boys. This is to be expected given the existence of single-sex schools in the data. Thus we look at only co-educational schools in the last two rows. Here, while the average female share is below 50% for both girls and boys, the average girl has 46% female share in her cohort, while the average boy has

only 39% females in his cohort.

In Figures 3.6a, 3.6b and 3.6c, we plot the gender wise take-up rate of Mathematics, PCM and Biology, respectively, for each decile of female share in cohort in co-educational schools. For Mathematics and PCM, we see that the take-up rate of boys is almost invariant to the share of girls in the cohort till the 5^{th} decile and turns slightly negative beyond that. The relationship is mildly positive for girls. After the 8^{th} decile, however, the take-up rate of Mathematics and PCM decidedly falls for both genders. More than 80% share of girls in the cohort in class X is, in general, negatively associated with choosing Mathematics and PCM in class XI.

For Biology, the curves are U-shaped. In the last decile of female share in cohort, the take-up rate of girls falls sharply, but that of boys shows increment even in the last decile. Thus, a very high concentration of girls in the cohort is positively associated with boys choosing Biology, but not girls. Finally, the graphs also plot the average take-up of these subjects in single-sex schools. In all cases, the take-up rate of these subjects is lower in both all-girls and all-boys schools, compared to co-educational schools. Thus presence of both genders is positively associated with the take-up of our three chosen subjects.

We now consider the second aspect of cohort peers: peer performance[15]. Here too we look at class X cohort peers of students and their performance in class X board examination. Panel A of Table 3.6 reports mean peer performance figures for all schools in the data. The peer set of girls are lower performing than that of boys. An average peer of a girl scores 345 out of 500 in class X, while that of a boy scores

[15]Peer performance is calculated by subtracting a student's own score from the sum total of cohort score of the school and then dividing the difference by the number of peers (total number of students in the school minus 1). Gender wise peer performance indicators are calculated similarly, except that the student's own score is subtracted only when own gender matches the gender of the peer group being considered.

352 on average. The next two rows depict performance figures for female and male peers separately. Again, boys have higher performing female peers than girls but slightly lower performing male peers than girls.

However, looking at only male and female cohort peers at a time eliminates single-sex schools for at least one gender[16]. Thus, we report the numbers for co-educational schools only in Panel B. Here we see that girls have an overall lower performing peer set a lower performing female peer set as well as a lower performing male peer set compared to boys.

In Figures 3.7a to 3.7i, we plot the take-up rate of Mathematics, PCM and Biology in each decile of cohort peer performance. The first row has overall peer performance deciles on the x-axis, the second row has female peer performance deciles and the last row has male peer performance deciles. For the last two rows, we only consider the sample of co-educational schools for comparability across genders. In Figure 3.7a and 3.7b, we see that the take-up of Mathematics and PCM increases with increase in peer performance for both boys and girls. Also, the gender gap exists almost unchanged at all deciles. In case of Biology, Figure 3.7c shows that whereas the take-up increases for both boys and girls with increasing peer score, the curve is much flatter for boys, and thus the gender gap also widens with increasing peer performance.

Next, in Figures 3.7d and 3.7e we see, upon closer inspection, that the gender gap for Mathematics and PCM starts wider among the lowest deciles and narrows down as we move up the female peer achievement deciles. The last decile sees a sharper decline in the gap mainly because the curve of girls bends up. Thus very high achieving female peers, and peers in general (Figures 3.7a and 3.7b), push up

[16]For example, looking at only female cohort peers eliminates only-boys schools and vice versa when looking at only male cohort peers.

the take-up rate of girls even above the mean rate for boys. For Biology, Figure 3.7f shows that the gender gap is actually reversed in the first decile of female peer performance: boys choose Biology more than girls. Then the relationship flips and, while the take-up rate increases monotonically for girls, the curve for boys is almost flat, with little movement away from the mean.

The third row of figures reveals that the pattern of take-up of the three subjects over the male peer performance distribution is different from what was observed in the previous row. For Mathematics and PCM where take-up is higher for boys, the gap is very close to zero in the first two deciles of male peer performance. Upon closer inspection, one can see that the take-up rate for girls is very similar to that observed in the second row in the lowest two deciles. It is the take-up rate of boys that has fallen. Lower performing male peers in the cohort pushes down the take-up rate of boys for Mathematics and PCM. After the bottom two deciles, the gap increases and stabilizes. For Biology, the gap is never zero, but is lower in the first two deciles. Thereafter, it widens are stabilizes.

Summing up, while the relationship between female share in cohort peers and stream choice is weak, a very high concentration of girls in cohort is negatively correlated with choosing Mathematics and PCM for both boys and girls. It is also negatively correlated with choosing Biology for girls, but positively correlated for boys choosing Biology. We also observe that while, in general, a positive relationship exists between the take-up rate of all the three subjects and the performance of cohort peers, the relationship between the gender gap in take-up and peer performance is much more varied. Girls appear to be closing some of the gap in the presence of very high achieving female peers in the case of Mathematics and PCM. Boys, on the other hand, are much less responsive to achievement of peers and thus, the gap in

Biology widens with increasing performance of cohort peers.

3.4.4 Immediate Seniors

In this section, we build on the premise that a student interacts with a number of people in a school environment. This includes her teachers, cohort peers as well as school seniors. Under this subsection, we use a student's immediate seniors when in class X to elicit two explanatory factors proposed in the literature. First is role models. It has been argued that students become less inclined to pursue fields in which there are no role models to emulate. Multiple studies have shown that instructors of the same gender, race or ethnicity can serve as role models for students and encourage them to enroll in streams they are underrepresented in (Bettinger and Long 2005; Hoffmann and Oreopoulos 2009; Paredes 2014; Fairlie et al. 2014; Bottia et al. 2015). However, students can also look up to their seniors as role models. Girls, for example, can feel more encouraged to opt for Mathematics or PCM after matriculation if more of their female seniors also opted for these subjects. Therefore, we consider how stream choices of seniors correlate with those of the juniors.

We have three consecutive cohorts in our dataset: class XII batches of 2014, 2015 and 2016 (who were in class X in 2012, 2013 and 2014, respectively). The class X of 2012 (2014 batch of class XII) is the immediate senior of the class of 2013, and the class X of 2013 is the immediate senior of the class of 2014. As we do not have data on class X batch of 2011, we only consider the later two cohorts for this part of the analysis, and use the two older cohorts for constructing our independent variables involving immediate seniors. Since we are looking at decisions made after class X, we consider seniors in the schools students were enrolled in class X.

To construct variables that measure the role model aspect of seniors, we calculate

the proportion of seniors of a student's own gender who opted for a given subject in *their* class XI. If g out of G girls in a girl's immediate senior cohort and b out of B boys in a boy's immediate senior cohort had chosen Mathematics, then g/G gives the role model measure for the girl student and b/B gives the corresponding measure for the boy student. We consider the sample of co-educational schools for comparability across genders. First, Table 3.7 reports the summary statistics of these variables. The first column reports that an average student has around 52% of own gender seniors who chose Mathematics, 43% who chose PCM and 19% who chose Biology. The next two columns shows these values for an average girl and an average boy student. As is obvious from the gender gap observed in stream choice, girls see only 41% of senior girls choosing Mathematics, while boys see 60% of senior boys choosing Mathematics. Similarly, girls see 20 percentage points lesser own gender seniors choosing PCM than boys. On the other hand, they see almost 28% seniors of their own gender take up Biology while boys see only 12% senior boys choose Biology.

The differences across genders in these role model measures are quite large and are also in line with the overall gender gap in the take-up of these subjects. If they also correlate positively with students' own stream choices then they can potentially explain some of the gender gap. We visualize how the stream choice of seniors correlate with students' own stream choices in Figure 3.8. Each graph in the figure has deciles of the share of own gender seniors choosing a subject on the x-axis. On the y-axis are the share of students in each of these deciles who themselves choose those subjects.

Our first observation is that there is a positive relationship between the share of seniors choosing a subject and the that of students who themselves choose that

subject in all three cases. Secondly, the gender gaps in all three subjects keeps reducing with increasing values of the role model measure. In fact, the gaps become zero by the last decile. From these graphs, it appears that seniors do serve as role models for students when making their own stream choice decisions and therefore, more own gender seniors in a subject can be associated with a higher take-up of that subject.

The second factor we examine using seniors is that of "Chilly Climate" (Clark Blickenstaff 2005). A number of studies have documented that when women enter male dominated fields of study or occupation, they encounter a hostile or an unwelcome environment. A male dominated field translates into overt or covert discrimination, a feeling of misfit or being at a loss (Sadker and Sadker 1986; Fouad et al. 2011; Lordan and Pischke 2016; Tellhed et al. 2017; Wu 2017). As mentioned earlier, this is even a possibility for males entering female dominated fields. We propose that students can gather an idea about what it would be like to enter into streams dominated by the other gender by looking at their seniors. In particular, students in class X can look at the gender composition of the senior Mathematics, PCM and Biology class and form a perception about how "chilly" the climate would be if they do choose these subjects post matriculation. If they see a higher share of their own gender among seniors who opted for the respective subjects, they may be more encouraged to opt for them. Out of Y seniors of a student who opted for a subject, if g are girls and b $(= Y - g)$ are boys, then g/Y gives the chilly climate measure for girls and b/Y gives the value for boys. The lower the value of this measure, the "chillier" the expected climate will be for a student.

Table 3.8 shows how the distribution of our measure of chilly climate varies across genders. Again, we only look at co-educational schools where students have both

male and female seniors. Column 2 reports the overall mean share of own gender students in senior Mathematics, PCM and Biology classes. On an average, a student sees that 55% of the seniors who opted for Mathematics or PCM are of her own gender, while 50% of Biology opting seniors are of her own gender. The next two columns report these values for girls and boys separately. Here, we see that while an average boy sees that almost 70% of the students in the senior Mathematics class are boys, an average girl sees that only 35% of those students are girls. The difference is even larger for PCM. The pattern expectedly flips for Biology with girls having 63% of own gender students in senior Biology class and boys having only 40%.

Finally, Figure 3.9 depicts the relationship between the gender composition of senior class and own subject choice. On the x-axes in each of the graphs are deciles of the share of own gender students in the senior subject class. The y-axes have the share of girls and boys who themselves opt for that subject after their class X. First thing to note is that all curves are inverted U-shaped, more so for the non-dominant gender in the subject. From Figure 3.9a we see that the take-up of Mathematics first increases with an increase in the share of own gender seniors in Mathematics class for both girls and boys, and goes over their respective means at some point. After that, the probability falls for both but the fall is very sharp for girls, going much below their mean levels. Figure 3.9b shows the same pattern for PCM.

In case of Biology, Figure 3.9c shows that the curve for girls follows a pattern similar to boys in the previous graphs, with a much wider range where the take-up is above the mean. It falls after the own gender share in senior Biology class crosses the 9^{th}. The curve for boys increases monotonically for the most part of the curve, goes above the mean at around the 4^{th}. It starts falling after around 8^{th}, but never goes below the mean.

In summary, this section indicates that immediate school seniors can act as gender role models for students and can potentially explain a good portion of the observed gender gap. The implications for the "Chilly" climate aspect of seniors, however, is different between boys and girls. While the relationship is overall positive for boys in all three subjects, it is positive for girls only at lower values of own gender share in senior subject classes.

3.4.5 Socioeconomic characteristics

Finally, in this subsection, we describe the background socioeconomic characteristics of students and also briefly discuss how they correlate with their stream choices. We already saw the statistics related to these variables in Table 3.1. Table 3.9 shows their distribution across girls and boys. The caste rows show that more girls than boys belong to General caste category, while more boys belong to the OBC category. Next we see that while fewer girls than boys are enrolled in Independent schools, more girls are enrolled in government schools. Finally, fewer girls than boys took the external board exam in their class X and lesser of them are single children in their families. The difference in the annual family income is not statistically significant due to the large standard deviations.

Table 3.10 shows how these attributes relate to the stream choices of students. From the first two panels, we see that Other Backward Classes or OBCs have the highest take-up of Mathematics and PCM overall, as well as for girls and boys separately, followed by the General caste. OBCs also have the highest take-up rate of Biology among girls, while Scheduled tribes (ST) have the highest take-up among boys. We also see that students who took the external board exam in class X have a higher than average rate (shown in Table 3.2) of choosing these subjects. The same

is the case with students who are the only children in their families. The most important thing to note, however, is that the gender gap exists in each of these categories, though the magnitude differs.

To summarize, in this section we have provided a detailed idea of what our explanatory factors mean, what variables are used to measure them, how they are distributed between girls and boys and how they are correlated with their stream choices and the gender gap in those choices. We are now equipped for the next section where we use regression and decomposition tools to econometrically dissect this gender gap and quantify how much of it can be accounted for by these explanatory factors.

3.5 Decomposition Analysis

Our aim in this section is to formally decompose the gender gap in the rate of take-up of Mathematics, PCM and Biology in our dataset. For this, we use a regression framework and a decomposition framework. Below, we a give general description of both the methodologies before we put them to use for each of our explanatory factor categories.

3.5.1 Regression framework

We run the following linear probability model:

$$S_{isc} = \beta_0 + \beta_1.F_{isc} + \beta_2.EF_{isc} + C_c + sch_s + \varepsilon_{isc}. \tag{3.1}$$

Here S_{isc} is the subject choice of student i in school s in cohort c. It is a binary variable which takes value 1 if the student chose subject S after class X,

and 0 otherwise. F_{isc} is a dummy variable for a female student. EF_{isc} stands for Explanatory Factor measured at the level of a student i in school s in cohort c. It can be socioeconomic characteristics of students, a measure of their ability, variables related to cohort peers or those measuring role model and chilly climate aspect of immediate seniors. C_c are cohort fixed effects to control for unobservables within a cohort and sch_s are school fixed effects to control for unobservables at the school level.

In this framework, we will monitor the movement of the female dummy variable, F_{isc}. Without any controls, the female dummy captures the raw difference between the take-up of a subject between boys and girls. We then add explanatory factors to see if the gendered pattern in stream choice is sensitive to the addition of these controls.

3.5.2 Decomposition framework

Pioneered by Blinder (1973) and Oaxaca (1973) to study wage gaps, decomposition techniques "decompose" a gap in a distributional statistic between two groups into an "explained" and a residual, "unexplained" component (Fortin et al. 2011). The explained component is the one due to a difference is endowments between the two groups. The unexplained part is due to a difference in the returns to those endowments. In this chapter, we use the Blinder-Oaxaca decomposition technique[17] (Jann 2008). Below is a brief discussion of the Blinder-Oaxaca decomposition methodology.

The Blinder-Oaxaca (BO) technique is a parametric decomposition method that assumes a linear relationship between the dependent and independent variables. It

[17]Since our dependent variable is binary, we also compute decomposition results using the Fairlie decomposition technique (not reported here) which uses a non-linear logit regression model (Fairlie 1999, 2005; Jann 2006). Our results are broadly similar using both the methodologies.

decomposes the gap in the means of the dependent variable between two groups using a linear regression model. Assume that B and G are the two groups and S is the outcome variable. Then the mean outcome difference, $D_O = E(S_B) - E(S_G)$, is decomposed into explained and unexplained components. Consider the linear model

$$S_z = X_z'.\beta_z + \varepsilon_z,\ E(\varepsilon_z) = 0,\ z\epsilon\{B, G\}, \tag{3.2}$$

where X is a vector of covariates and a constant, β is a vector of slope coefficients and the intercept term, and ε is the error term. The mean difference, D_O, can be written as

$$D_O = E(S_B) - E(S_G) = E(X_B)'\beta_B - E(X_G)'\beta_G, \tag{3.3}$$

using the assumption in (3.2) (Jann 2008). Now adding and subtracting $E(X_G)'\beta_B$ to (3.3) and rearranging, we get

$$D_O = E(S_B) - E(S_G) = \{[E(X_B) - E(X_G)]'\beta_B\} + \{E(X_G)'[\beta_B - \beta_G]\}. \tag{3.4}$$

The first term in curly brackets is the explained component of the gap or difference, D_O, explained by differences in observable covariates between the groups. The second term is the unexplained component of D_O: unexplained by observable differences in endowments, explained rather by a difference in the returns to those endowments. An intuitive way to understand a decomposition is to think of a counterfactual situation. In our context, for instance, if ability "explains" χ% of the gender gap D_O, it implies that if girls had the mean ability level of boys, and the return to ability like boys (β_B), then the gap D_O would be reduced by χ%. In other words, χ% of the gap is due to a difference in the mean ability levels of girls and boys.

It is commonplace in the decomposition literature to use the coefficients of the dominant group to calculate the explained part of the decomposition in the wage gap. The dominant group, in general, has higher returns to the given set of attributes like education and experience. This is not always the case in our context, as we will see in the next section. We instead use coefficients from a pooled model to determine the explained component as proposed by Neumark (1988) and Oaxaca and Ransom (1994). To avoid overestimation of the explained part due to inappropriate spillage of some of the unexplained component into the explained component, we also include a dummy for a female student as an additional covariate in the pooled model (Jann 2008). Thus, the explained component becomes $\{[E(X_B) - E(X_G)]'\beta_P\}$, where β_P is the coefficient from the pooled regression and is assumed to be the same for girls and boys.

We now delve into each of the categories of explanatory factors one by one and report the results in the next section.

3.6 Results

3.6.1 Socioeconomic characteristics

We begin this section with reporting the results with socioeconomic characteristics as explanatory factors because we will use them as basic controls in all the subsequent analyses. Table 3.11 reports the regression results. All regressions include only cohort fixed effects to begin with and add class X school fixed effects in the last column for each subject. We cluster the standard errors at the class X school level.

Columns 1 to 3 show results for Mathematics. In column 1 of Table 3.11 we see that with only cohort fixed effects, the coefficient of the female dummy is -0.191

which is almost equal to the raw gender gap in the take-up of Mathematics we saw in Table 3.2. Column 2 adds socioeconomic status (SES) variables, which includes dummies for caste, (General caste is the omitted category), annual family income, single child status and a dummy for taking external board exam in class X as opposed to school based board exam. The inclusion of these variables reduces the size of the female dummy coefficient by 0.8 percentage points. We also see that while OBCs have a statistically significantly higher probability of choosing Mathematics after class X, both SCs and STs have a statistically significantly lower probability than the General caste students. This is in line with our observations from Table 3.10. The coefficient on family income, while statistically significant at the 10% level, is negligibly small in magnitude. The coefficient is statistically insignificant for the single child dummy, while it is positive and statistically significant for the external board exam dummy. In column 3 school fixed effects are added. The size of the female dummy coefficient falls further by 0.7 percentage points to -0.176. The signs and statistical significance of most of the other coefficients remain unchanged (except the single child dummy which now becomes statistically insignificant).

The results for PCM in columns 4 to 6 mirror those of Mathematics. The raw gender gap of 20.61 percentage points in Table 3.2 remains intact after the inclusion of cohort fixed effects in column 4. When SES variables are added in column 5, the coefficient of the female dummy falls by 0.9 percentage points. Here too, SC and ST have statistically significantly negative coefficients, while OBC has a statistically significant positive coefficient. Family income, single child status and external board exam dummy have positive and statistically significant coefficients. Finally school fixed effects are added in column 6. Similar to Mathematics, the size of the female dummy coefficient falls further by 0.6 percentage points while the other coefficients

have the same sign and significance (except annual family income which becomes statistically insignificant).

The last 3 columns look at Biology as the subject choice. The raw difference of 11.61 percentage points observed in Table 3.2 increases slightly to 11.7 percentage points when cohort fixed effects are added. The addition of SES variables in column 8 leads to an increase in the gender gap to 12.1 percentage points. All the caste dummies added in column 8 have positive and significant coefficients: all of them have a higher probability of choosing Biology than the General caste. Family income has positive and statistically significant coefficient, but of a very small magnitude. Lastly, single child status and external board exam in class X both have positive and statistically significant coefficients. Adding school fixed effects in the last column further increases the coefficient of the female dummy by 3.7 percentage points so that the final gap stands at 15.8 percentage points after controlling for cohort fixed effects, school fixed effects and SES variables.

To summarize, the socioeconomic status variables reduce the size of the female dummy coefficient for Mathematics and PCM, but not for Biology. School fixed effects reduce the size of this coefficient further for Mathematics and PCM and increases it further for Biology. We include the variables in Table 3.11 as base level controls (along with cohort and school fixed effects) in the analysis of all the other explanatory factors.

3.6.2 Ability

We now add student ability to our decomposition analysis using the variables described in Section 3.4.2. Table 3.12 reports results from the linear probability models. Columns 1 and 2 has Mathematics as the subject choice, columns 3 and 4 have PCM

and columns 5 and 6 have Biology as the subject choice. The first columns for each subject reproduce the respective last columns from Table 3.11 which we use as our starting point for each subject and add ability variables to them. These include the class X total score, class X Mathematics and Science scores and comparative advantage terms[18]. To recall, a student's comparative advantage in subject A vs. B is measured as the difference in her standardized scores in these two subjects.

For Mathematics, we see from column 2 of Table 3.12 that class X total score has a negative and statistically significant coefficient. While this may seem counter-intuitive, we must note that the specification includes all the other ability terms and the *net* return to class X total score is negative. The size of the coefficient, however, is very small. Next, we see that the coefficients on both Mathematics score and Science score are positive and statistically significant. Finally, a comparative advantage in Mathematics vs. English has a large, positive and statistically significant association with choosing Mathematics after class X, while a comparative advantage in Science vs. Mathematics has a negative and statistically significant correlation. Importantly, addition of various measures of absolute and comparative advantage in ability actually increases the magnitude of the female dummy to almost the raw gender gap of 19.13 percentage points.

Columns 3 and 4 show very similar results for PCM. All terms, except the class X Mathematics score, have coefficients that are similar in magnitude, sign and statistical significance to those of Mathematics. The coefficient of the class X Mathematics score has a statistically insignificant coefficient. Addition of all ability terms together

[18]Since adding all three comparative advantage terms together leads to collinearity and one of the terms gets dropped, we include only two of them for each subject choice. A comparative advantage in Mathematics vs. English and in Science vs. English are added for Mathematics and PCM while a comparative advantage in Science vs. English and in Science vs. Mathematics are added for Biology.

increases the magnitude of the female dummy close to the raw gender gap for PCM too.

Columns 5 and 6 have Biology as the subject choice. Here class X total score has a positive and statistically significant coefficient, as does the coefficient of class X Mathematics score. Class X Science score has a negative and statistically significant association with the take-up of Biology. Lastly, a comparative advantage in Science vs. Mathematics has a large and positive coefficient which is statistically significant. The coefficient of the Science vs. English comparative advantage term has a negative coefficient. Most importantly, the ability terms purge the size of the female dummy coefficient by 2.3 percentage points. Thus we can expect the ability related attributes to explain some significant portion of the gender gap in Biology.

Table 3.13 reports results from the Blinder-Oaxaca decomposition. We measure the gap as take-up of boys minus that of girls. As explained in Section 3.5.2, the coefficients from Table 3.12 will be used to calculate the explained components. Columns 1 and 2 report results for Mathematics, followed by PCM in columns 3 and 4 and Biology in columns 5 and 6. The first row gives the total gap in the take-up in percentage points. The following rows give the contribution of SES and ability variables in percentage points. The numbers in the brackets below give the contribution as a percentage of the total gap ((percentage points explained/total gap)*100). The odd numbered columns only have the socioeconomic status variables. First consider Mathematics and PCM. The SES variables explain 1.52% of the total gap of 19.08 percentage points in Mathematics and 1.65% of the 20.57 percentage points gap in PCM. The ability variables are added in the next columns. Class X total score explains 1.15% of the gap in Mathematics and 1.65% of that in PCM. The contribution of the class X Mathematics score is statistically insignificant for

both the subjects. The contribution of the class X Science score, on the other hand, is negative. To understand this mathematically, recall that the "explained" part of the gender gap is given by the first term of equation (3.4): $\{[E(X_B) - E(X_G)]'\beta_P\}$, where β_P is the coefficient from the pooled regression of Table 3.12. Since girls have a higher mean Science score than boys, the expression inside the brackets is negative in sign, while β_P is positive, resulting in a negative product. Intuitively, this can be understood by using the counterfactual exercise. If girls had the mean Science scores of boys, which is lower than their actual scores, then the gender gap would, in fact, be higher than what is observed, since Science scores are positively correlated with the probability of choosing Mathematics.

The largest contribution to the explained part comes from the gender differences in comparative advantage in Mathematics vs English. It explains around 15% of the gap for both Mathematics and PCM. The Science vs. English comparative advantage term, however, has a negative contribution to the explained part. This is to be expected given the negative coefficient of this term in Table 3.12 and that girls have a higher mean value of this attribute. The last row shows that all the ability terms together explain 7.44% of the gender gap in Mathematics take-up and 5.54% of that in PCM take-up.

For Biology, the total gap is -11.65 percentage points because boys are the reference group and their take-up rate of Biology is lower than girls. Column 5 shows that when only SES variables are added, the explained percentage is negative. This means the predicted gap is higher: if girls had the SES levels of boys, they would have even higher probability of choosing Biology. Thus the gender gap will be larger. Class X total score explains 4.89% of the gap. Boys have lower class X total score on average, and if girls had scores like boys, they would be lesser likely to choose Biology and

thus the gender gap will be smaller. The contributions of class X Mathematics and Science scores are statistically insignificant. For Biology too, the largest contribution comes from the comparative advantage term. A comparative advantage in Science vs. Mathematics explains almost 10% of the gender gap in Biology. If girls were counterfactually given lower mean values of Science vs. Mathematics comparative advantage, they would be 9.53% less likely to choose Biology. All the ability factors together explain 9.10% of the 11.65 percentage points gender gap in Biology.

Summing up, comparative advantage in the relevant subjects turns out to be the largest explanators under the ability head. It explains around 15% of the gap in Mathematics and PCM and around 10% of the gap in Biology. Taken together, ability differences can explain under 10% of the gap in the three subjects.

3.6.3 Cohort Peers

We now consider our next set of explanatory factor: cohort peers. Table 3.14 reports the results from the linear probability models. As before, columns 1 and 2 have Mathematics as the dependent variable, followed by PCM in columns 3 and 4 and Biology in the last two columns. All specifications have cohort and school fixed effects as controls. As our base regressions, we reproduce equivalents of columns 2, 4 and 6 of Table 3.11 in the odd numbered columns for the sample co-educational schools. To them we now add peer composition and peer performance variables discussed in Section 3.4.3. The even numbered columns add measures of female share in cohort, average peer score and gender-wise average peer scores[19].

For all three subjects, we see that female share in class has a positive and sta-

[19]The peer score variable is calculated after subtracting own score from the sum of scores of all students in a school-cohort and then dividing the difference by the total number of students in the school-cohort minus one.

tistically significant coefficient. Average peer score in class X has negative and statistically insignificant coefficients for Mathematics and PCM, but negative and statistically significant coefficient for Biology. The magnitude, however, is small for all three cases. Finally, we see that the average scores of both male and female peers have small, positive and statistically significant coefficients for all the subjects. Most important thing to note, however, is that the addition of all the peer related terms does not change the coefficient of the female dummy in any way for any of the subjects.

Table 3.15 shows the decomposition results. Column 1 has Mathematics as the subject choice, column 2 has PCM and column 3 has Biology. The first row shows the total gender gap in this sample of students from co-educational schools.

The first and the most important observation from Table 3.15 is that the contributions of none of the peer related variables are statistically significant in any of the columns for any of the subjects. Additionally, those contributions also have very small magnitudes. Female share in cohort has negative contributions for the gender gap in Mathematics and PCM and positive contribution for Biology. Same is true for the average peer score term. The gender wise peer terms have positive contributions for Mathematics and PCM and negative contributions for Biology. However, none of these contributions are statistically significant. This is because the distribution of the attributes are not very different across boys and girls so that assigning the attributes of boys to girls does not lead to statistically significant changes in the take-up rate of girls. We can conclude this subsection with the takeaway that differences in peer related variables do not contribute to explaining any of the gender gap in the take-up of Mathematics, PCM and Biology.

3.6.4 Immediate Seniors

We now move on to consider our last set of explanatory factors where we use a student's immediate seniors to elicit the role model and chilly environment aspects of making a stream choice. As explained in Section 3.4.4, only the two younger cohorts can be used for this part of the analysis. The two older cohorts – class X batches of 2012 and 2013, are the immediate seniors of the two younger cohorts – class X batches of 2013 and 2014, respectively.

Table 3.16 reports results from the linear probability models. We again start with the specification with SES controls, cohort fixed effects and school fixed effects and run it on the sample of co-educational schools in the later two cohorts to get our base regressions. To this we add senior related variables in the even numbered columns: the share of own gender seniors who opted for the subject (role model measure) and the share of students of own gender in the senior Mathematics, PCM or Biology class (measure of chilly climate).

In Table 3.16 we see the biggest reductions so far in the size of the coefficient of the female dummy. Addition of our measures of senior role models and expected chilliness in climate lowers the value of the female dummy coefficient by 5.9 percentage points for Mathematics, by 5.4 percentage points for PCM and by 2.9 percentage points for Biology. In addition, the coefficients on these terms are themselves always positive and statistically significant, except for that of the measure of chilly climate for Biology. These observations, along with the large differences in the distribution of these attributes across genders, suggest that they will be able to explain a significant part of the gender gap.

Table 3.17 shows that this is indeed the case. Again, column 1 presents the decomposition results for Mathematics, column 2 for PCM and column 3 for Biology.

We see that for Mathematics, our measure of chilly climate emerges as the largest explanator so far, explaining as much as 23.53% of the 18.74 percentage points gender gap. If girls had the much higher share of own gender students in the senior Mathematics class like boys, they would be almost 24% more likely to themselves opt for Mathematics after class X. For PCM too, it can explain almost 16% of the gap in take-up, slightly higher than what is explained by a comparative advantage in Mathematics vs. English (column 4, Table 3.13). Our measure of role model explains around 11% of the gap in PCM, while its contribution is statistically insignificant only for Mathematics.

For Biology, on the other hand, it is our measure of role model which explains the largest portion of the gender gap so far. If girls had the lower levels of the share of own gender seniors choosing Biology, the gender gap would be closed by almost 18%. The contribution of the chilly climate measure is statistically insignificant for Biology.

Overall, our senior related variables are able to explain 31% of the gender gap in Mathematics take-up, 26.78% of that in PCM take-up and 18.21% of the gap in Biology take-up. It is interesting to note that the expected chilliness in the climate of a course dominated by the opposite gender is the most important factor for girls for avoiding that course (compared to the other explanatory factors studied in this chapter). It is also interesting that this is not a factor for boys. Thus, while a skewed sex ratio discourages girls to enter in Mathematics and PCM, this is not an important factor for boys when making a choice to take up female dominated courses like Biology.

To sum up, the association between the subject choices of students and those of their immediate own gender seniors are strongly positive. In addition, boys and girls

have large differences in these senior related attributes. As a result, if girls counterfactually had own gender seniors like boys have, the gender gap in Mathematics, PCM and Biology would be lower by 20% to 30%. It is important to note that all the underlying regression models in the decomposition exercises control for cohort and school fixed effects. Thus, these differences do not reflect unobservables at the school level. They capture the correlation between the decision of a student and his/her seniors within a school.

3.7 Discussion

In this section, we briefly discuss our findings in the previous sections and their implications in the Indian context.

Our aim in this chapter is to establish both the direction and the magnitude of the gender gap in stream choices in India, and then to decompose this gap to understand the factors that can possibly account for this. In the first part of the analysis, we show that the stream choice in India follows the same gendered pattern as those in the developed countries. Using a rich student level dataset, we are also able to tease out the fine differences between mathematical and non-mathematical aspects within the Science stream. We find that boys are not more likely to opt for Science subjects in general. Rather, the gender divide exists at the level of mathematical rigour of a subject. While more boys choose Mathematics and PCM, more girls choose Biology. The gender gap in the other subjects most commonly chosen by students is, on the other hand, always below 10%. The largest gaps exist in subjects which are also the most economically rewarding.

In the next part, we begin to dissect this gap using the Blinder-Oaxaca decomposition technique. Our first step is to critically examine the explanatory power of

student ability towards the gender gap in Mathematics, PCM and Biology. The first set of our decomposition results clearly indicates that absolute ability, as measured by class X scores of students, explains only a negligible part of the gender gap in these subjects. However, a comparative advantage in the relevant subject can explain up to 15% of the gender gap. This result indicates that ability and aptitude differences are much more nuanced than the crude measures commonly used.

In the next set of results, we show that gender differences in peer attributes are small enough so that they can not explain any statistically significant part of the gender gap in stream choice. While the evidence in the literature suggests that girls and boys are differentially sensitive to peer composition and peer performance, we find that the peer set is not dramatically different between the genders in our dataset. Thus, counterfactually assigning the peer attributes of boys to girls will not change their stream choice decisions.

Lastly, we utilize a student's immediate seniors and their stream choices to form measures of her expectations regarding a subject choice. We postulate that own gender seniors can serve as a source of role models for students and that the gender composition of the senior subject class can help students form an expectation about how "chilly" the climate is going to be if they themselves choose a subject that is dominated by the other gender. We find that this source of gender differences can explain up to 30% of the observed gender gap in student stream choices. More importantly, we find that the results are asymmetric by whether a subject is dominated by girls or boys. In case of Mathematics and PCM, which are dominated by boys, we find that the expectation of a more chilly climate explains a large part of why girls are less likely to opt for these subjects. On the other hand, for a female dominated subject like Biology, this factor is not important. There, it is the lack of

own gender role model seniors which is the most important explanatory factor. Our results suggest that the fear of being surrounded by "too many" of the other gender is a concern for girls, but not for boys. It is important at this point to recognize that both chilly climate and role model effects could be driven by some deeper cultural or social norms, which result in girls and boys making similar choices cohort after cohort. However, the gender asymmetry in these results bolsters our claim that we are not picking up something entirely mechanical. Even though they have the same seniors, girls and boys take into account different aspects of their seniors' stream choices when making their own stream choices.

A caveat in our findings is that at least 69% of the gap remains unexplained even after accounting for all of our explanatory factors[20]. Below we discuss the possible sources for this residual gap and the scope of future research.

One of the major explanatory factors studied in the literature, which we were unable to examine given our dataset, is a difference is preferences across genders. A number of studies have shown that girls and boys have very different preferences regarding the non-pecuniary aspect of a career (Montmarquette et al. 2002; Baram-Tsabari and Yarden 2011; Kahn and Ginther 2017; Wang and Degol 2017; Patnaik et al. 2020). For example, girls place a higher value on enjoyability of the course material (Zafar 2013; Wiswall and Zafar 2015), people oriented careers (Diekman et al. 2010; Eccles and Wang 2016) and family-work life balance (Bronson 2014; Wasserman 2015) than boys. These differences could be a result of social conditioning where men and women are required to fill in preordained gender roles. Differential risk and confidence across genders could also shape their preferences. Finally, especially in the context of India, decision making agency is also crucial. Girls (and often boys)

[20]The highest percentage of the gender gap is explained for Mathematics in Table 3.17, which is 31%.

may not be choosing streams based on their preferences, but based on what their parents deem fit for them. The choice of parents could, in turn, be driven by marriage market demands, cost of education, willingness to spend on education as well as on availability of higher education institutes in their locality. For example, girls may not be allowed to take Mathematics or PCM because pursuing STEM programmes after high school either involves a higher cost of education, or a relevant college is not available in the vicinity and parents are reluctant to send their daughters to far away colleges. It could also be because a girl child is not expected to continue education after school. A boy, similarly, may be discouraged to opt for Biology because of social desirability bias, or even because a career in medicine has a long gestation period and boys are expected to start earning early on in their life.

The foremost implication of these is that the girl-boy difference in stream choice is not superficial. Rather, it is deep rooted in cultural upbringing and societal expectations (Giuliano 2020). The policies designed to bridge this gap have to incorporate these nuances and target deeper issues of gender stereotyping, marriage market functionalities, and post marriage balance of power and division of labour in the family. A most pressing future area of research would be to link career choices with marriage market variables. We conclude this section with the following caveat: To the extent that preferences are primitive, policies designed to bridge the gender gap in subject choices may not necessarily be welfare enhancing.

3.8 Conclusion

In this chapter we examine in detail the first stream choices of students in India after class X. Using three cohorts of results data from the Central Board of Secondary Education (CBSE), we first quantify the extent of the gender gap in subjects chosen

in higher secondary school. We find that there is a 19.13 percentage points difference in the take-up Mathematics between boys and girls. The gap is 20.61 percentage points in PCM choice. Boys are also 6 percentage points more likely to opt for computer Science. Girls, on the other hand, are 12.61 percentage points more likely to choose Biology and 5 percentage points more likely to choose Arts and Economics.

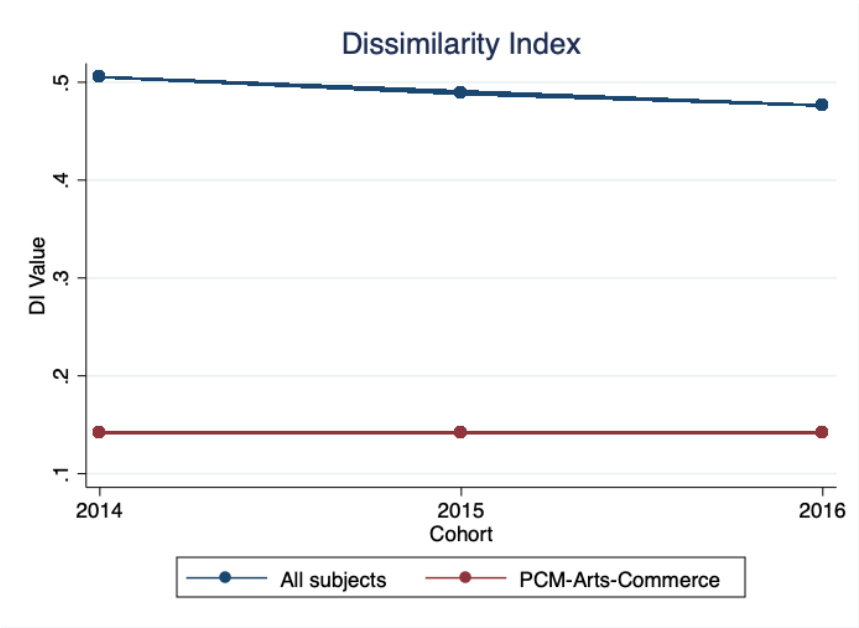
Next we describe each category of our explanatory factors one by one and illustrate graphically how they are related with stream choices of students. We also document how they are distributed differently across girls and boys. The largest differences between the genders are seen in the attributes related to their immediate seniors. Then we examine each explanatory factor under a regression framework and a decomposition framework. For the regression analysis, we use a linear probability model. We introduce the various categories of explanatory variables and see how they change the coefficient of the female dummy which gives the magnitude of the gender gap. Here too we find that role model and chilly climate aspects of immediate seniors bring about the largest reductions in the magnitude of the female dummy. A comparative advantage in the relevant subject turns out to be the second most important factor.

With this information, we finally put the explanatory factors in a decomposition framework one by one using the Blinder-Oaxaca linear decomposition technique. We report three broad findings. One, role model and chilly climate aspects of a student's immediate seniors in school are the largest explanators of the gender gap in Mathematics, PCM and Biology. If girls had the share of own gender students in the senior Mathematics and PCM classes like boys, and the share of own gender seniors choosing Biology like boys, the gender gap in these subjects would have closed by 24%, 16% and 18%, respectively. Two, for Mathematics and PCM, a

comparative advantage in Mathematics vs. English is the second largest contributor to explaining the gap, while it is a comparative advantage in Science vs. Mathematics for Biology. Three, peer composition and peer performance variables do not explain any statistically significant portion of the gender gap in any of the three subjects. We also note that even after accounting for gender differences in a wide range of attributes, at least 69% of the gender gap in our data remains unexplained. The differences observed in the form of stream choices actually masks huge differences in upbringing, expectations and balance of power between men and women from a very young age.

Figures and Tables for Chapter 3

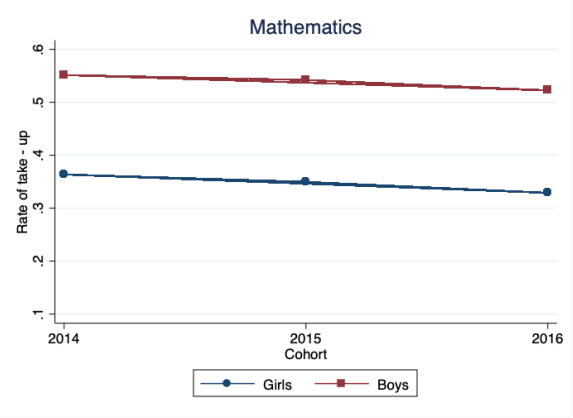
Figure 3.1: Dissimilarity Index



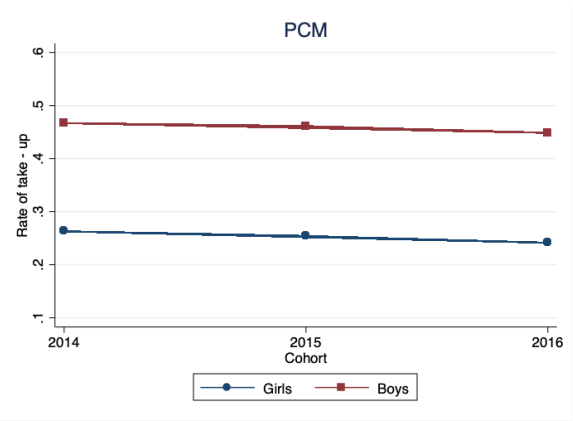
Note: The figure plots Dissimilarity Index (DI) for subject choice in the three cohorts in the data. DI captures the proportion of students who would need to change subjects in order for the distribution in each subject to look like the gender distribution in the entire sample. Data source: Central Board of Secondary Education.

Figure 3.2: Cohort-wise subject choice

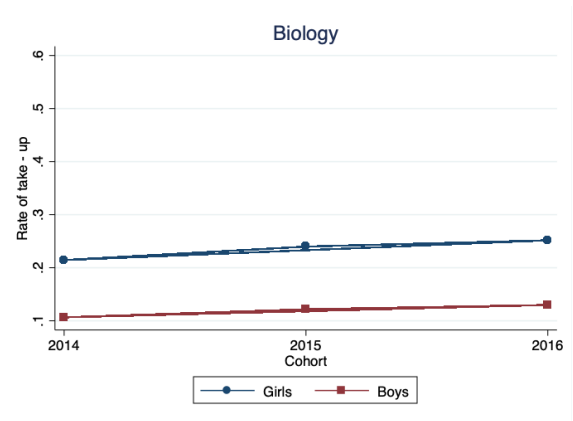
(a)



(b)

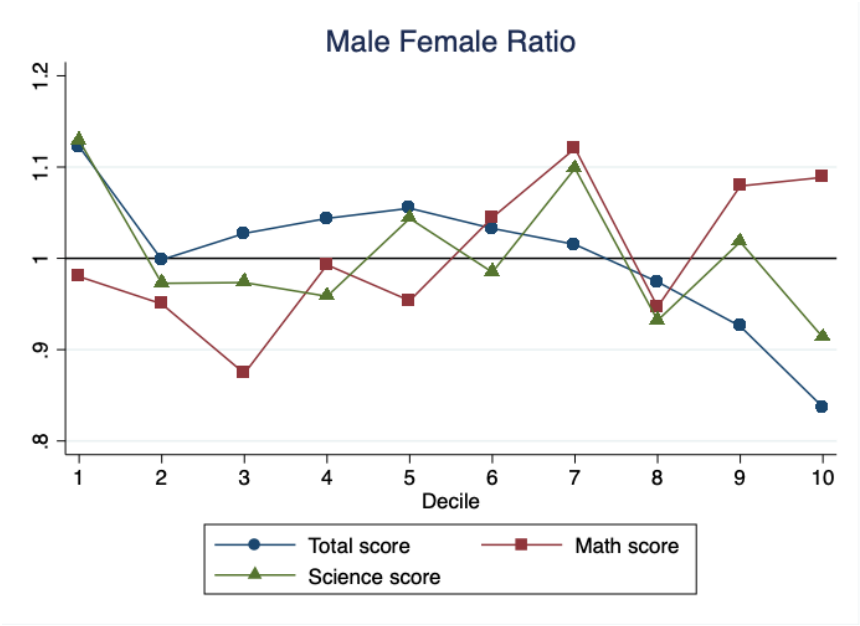


(c)



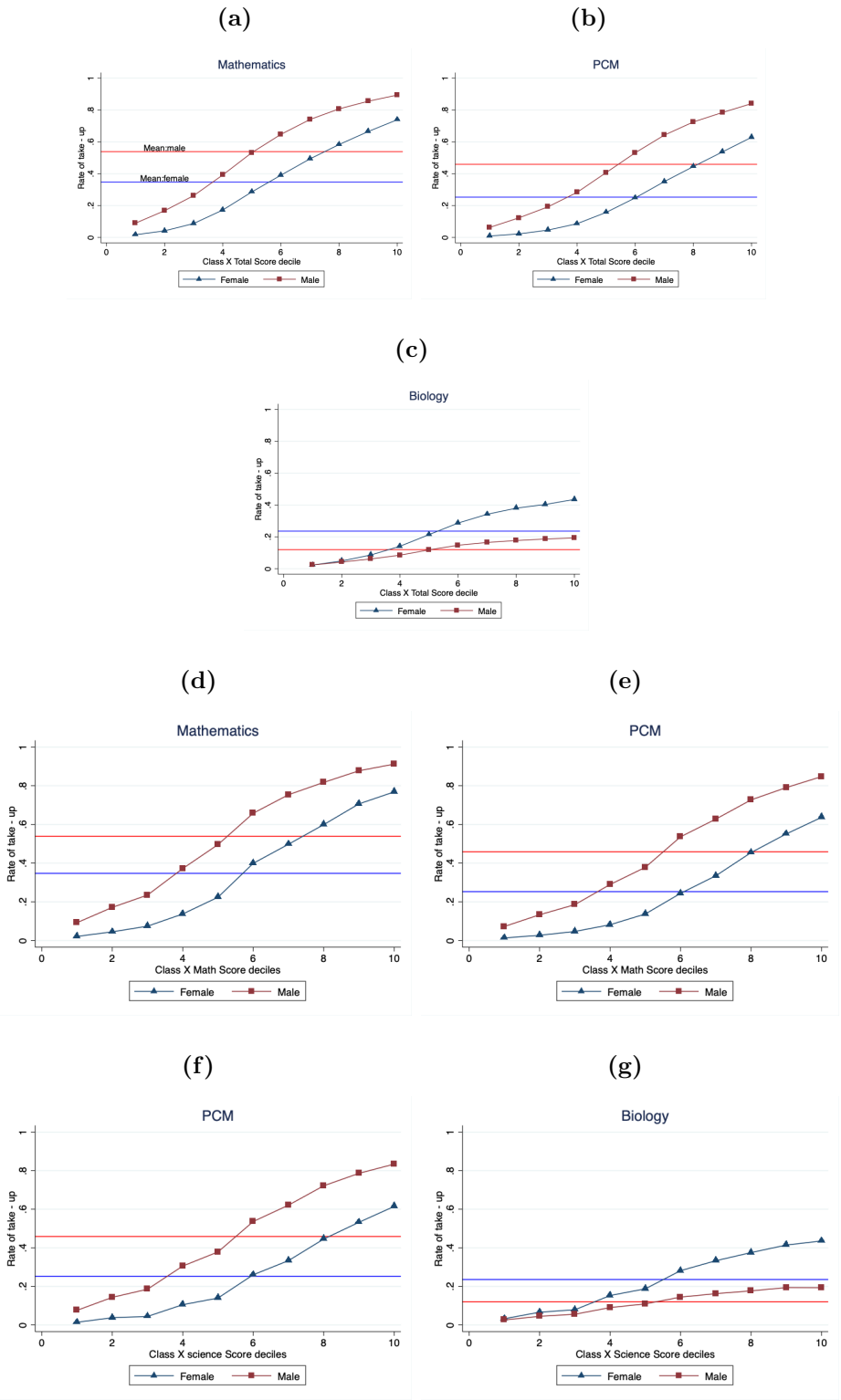
Note: The graphs plot the gender wise proportion of students who opt for of Mathematics, PCM and Biology after matriculation for each cohort in the data. Data source: Central Board of Secondary Education.

Figure 3.3: Boy-girl ratio across score deciles



Note: The graph plots the gender wise proportion of students over deciles of class X total and subject scores. Data source: Central Board of Secondary Education.

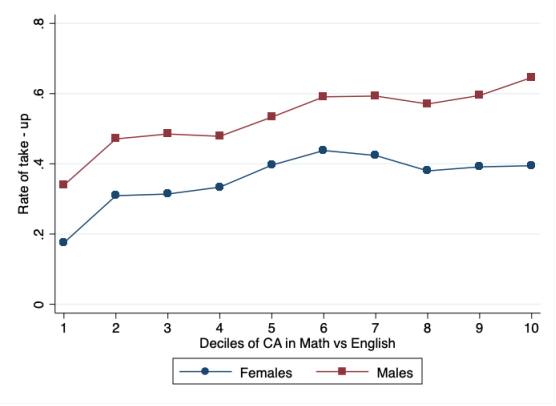
Figure 3.4: Stream choice by class X scores



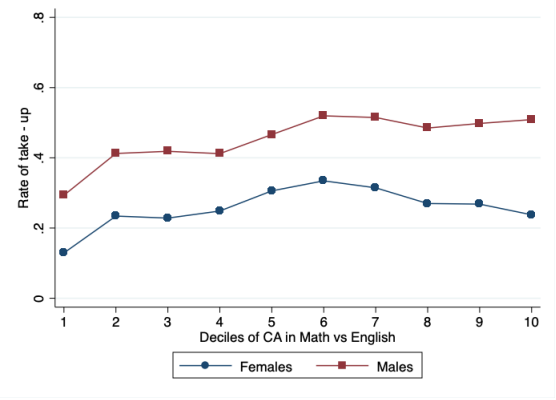
Note: Each graph plots the gender wise proportion of students who opt for a given subject after matriculation over class X score deciles. The first row plots the take-up of Mathematics, PCM and Biology over class X total score deciles. The second row plots the take-up of Mathematics and PCM over class X Mathematics score deciles. The third row plots the take-up of PCM and Biology over class X Science score deciles. Data source: Central Board of Secondary Education.

Figure 3.5: Comparative advantage and stream choice

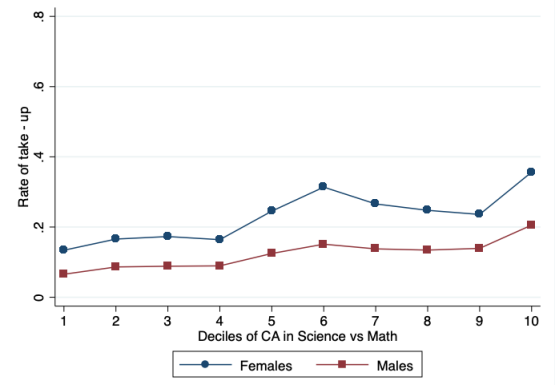
(a) Mathematics



(b) PCM



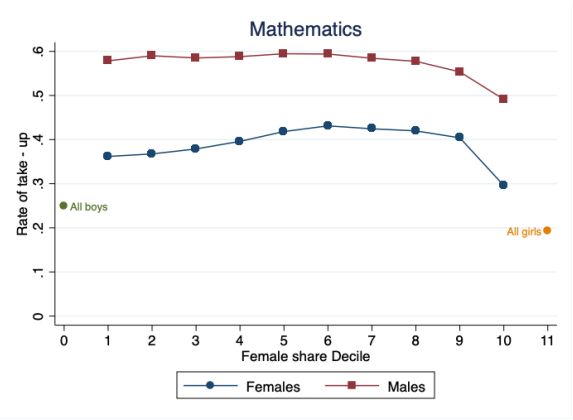
(c) Biology



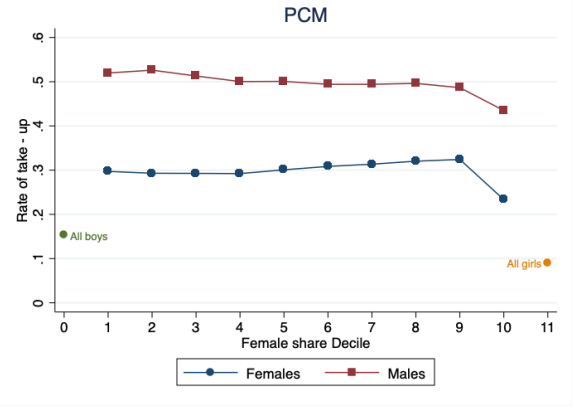
Note: The graphs plot the gender wise proportion of students who opt for Mathematics, PCM and Biology post matriculation over deciles of comparative advantage in relevant subjects. Comparative advantage in one subject over another is measured as the difference in their standardized scores in class X board examination. The first two graphs plot the rates of take-up of Mathematics and PCM, respectively, over deciles of comparative advantage in Mathematics vs English. The last graph plots the rate of take-up of Biology over deciles of comparative advantage in Science vs Mathematics. Data source: Central Board of Secondary Education.

Figure 3.6: Female share and stream choice

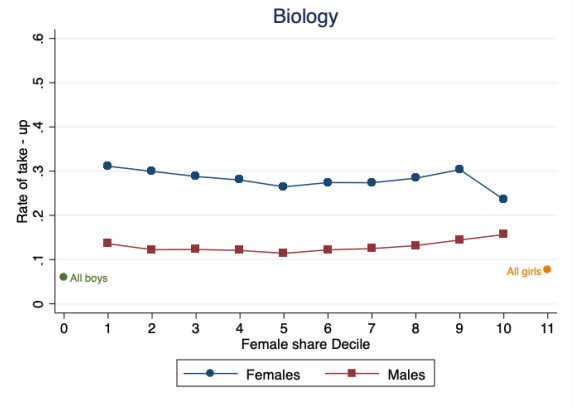
(a)



(b)

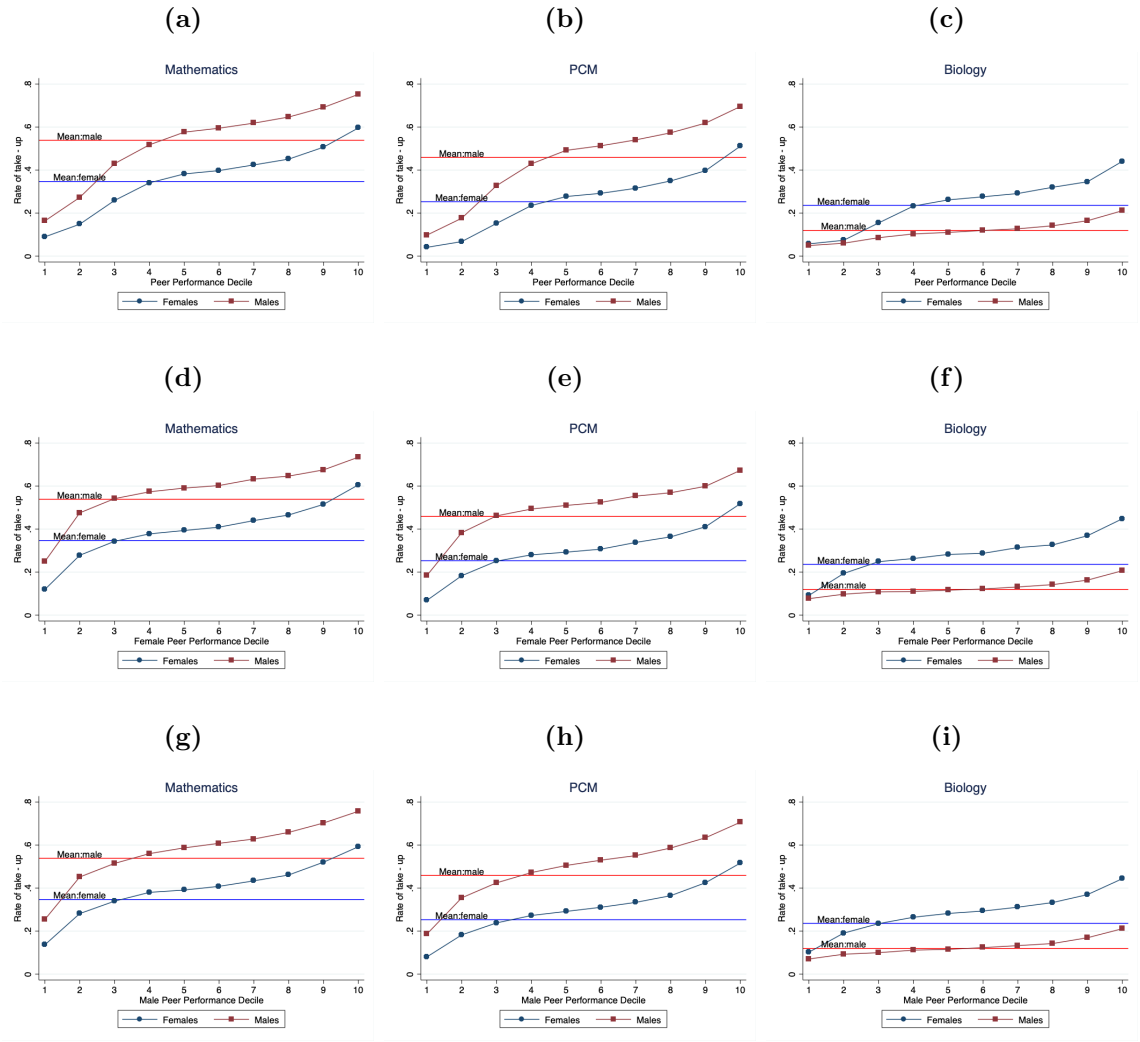


(c)



Note: The graphs plot the rates of take-up of Mathematics, PCM and Biology, respectively, post matriculation over deciles of female share in school in a cohort. Data source: Central Board of Secondary Education.

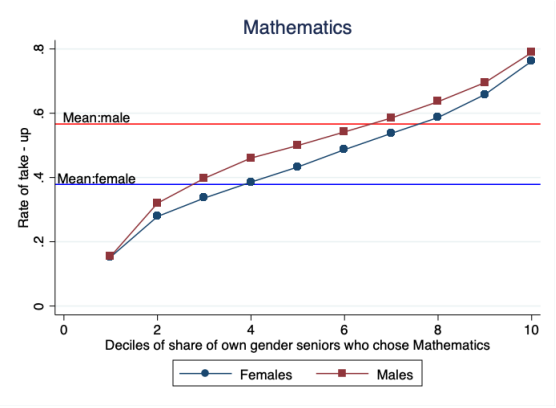
Figure 3.7: Peer performance and stream choice



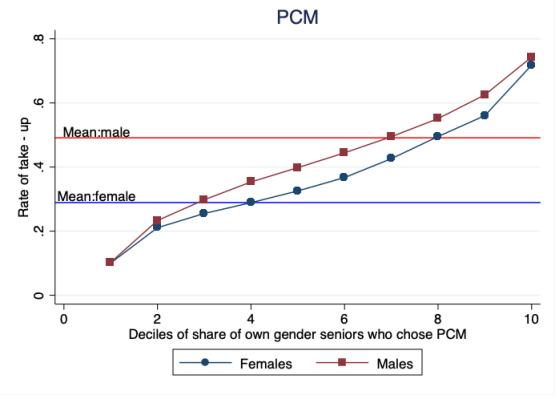
Note: Each graph here plots the gender wise proportion of students who opt for a given subject over deciles of mean class X scores of cohort peers in the school. Mean peer score is calculated by subtracting own score from the subtotal of scores of all students in the school in the cohort, and dividing the difference by the number of students in the school-cohort minus 1. The first row plots the take up Mathematics, PCM and Biology over deciles of scores of all peers. The second row plots the take-up rate of these subjects over score deciles of female peers and the last row plots them over score deciles of male peers. Sample is restricted to co-education schools in the last two rows for comparability across genders. Data source: Central Board of Secondary Education.

Figure 3.8: Seniors as role models and stream choice

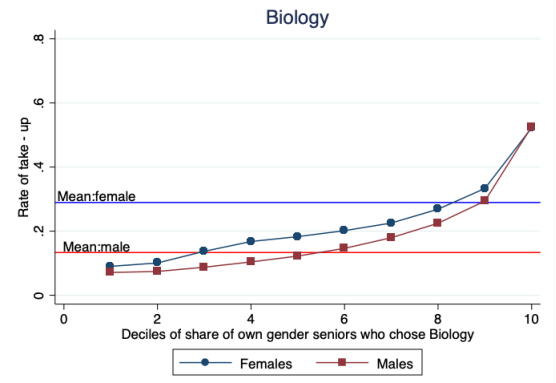
(a)



(b)



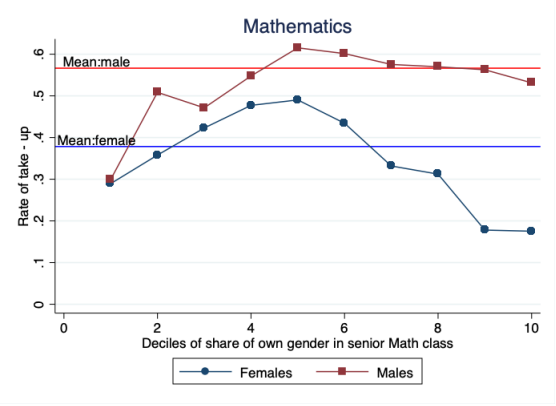
(c)



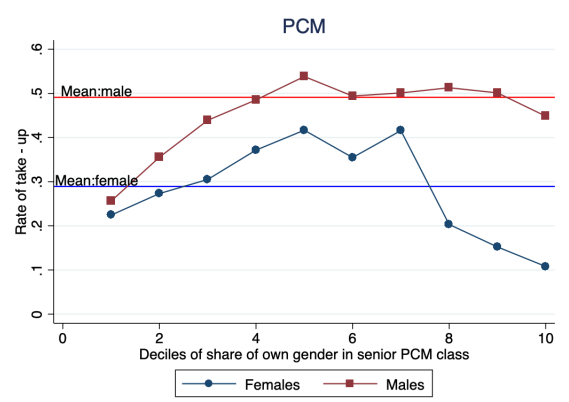
Note: The graphs plot gender wise proportion of students who opt for a given subject over deciles of the measure of role model. The role model variable is measured as the proportion of own gender seniors who opted for the subject. The first graph has Mathematics as the subject choice, followed by PCM in the second and Biology in the third graph. Data source: Central Board of Secondary Education.

Figure 3.9: Chilly climate and stream choice

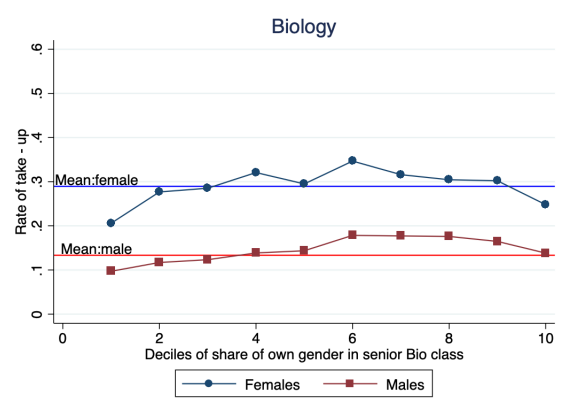
(a)



(b)



(c)



Note: The graphs plot gender wise proportion of students who opt for a given subject over deciles of the measure of chilly climate. The chilly climate variable is measured as the proportion of own gender students in the senior subject class. The top graph has Mathematics as the subject choice, followed by PCM in the second and Biology in the third graph. Data source: Central Board of Secondary Education.

Table 3.1: Summary Statistics

Total students	24,03,957
Girls	43.99%
Mean age in class X	16.68 years
Caste	
General	73.06%
Scheduled Castes	7.31%
Scheduled Tribes	3.33%
Other Backward Castes	16.31%
Type of school administration	
Private Aided	1.68%
Government	19.56%
Independent	68.03%
Jawahar Navodaya Vidyalaya	3.04%
Kendriya Vidyalaya	7.40%
Type of board exam in class X	
External board exam	37.55%
School based exam	62.45%
Mean score (out of 500)	
(std. dev)	
Class X	348.72
	(73.96)
Class XII	329.91
	(83.68)
Other attributes	
Mean Annual family income (INR)	292328.8
(std. dev)	$(17,44,665)$
Single child	5.78%
Disability	0.20%

Note: Castes are the administrative caste categories in India. Data source: Central Board of Secondary Education.

Table 3.2: Subject choice

Subject	Share of students (percent)	Of which girls (percent)	Share of girls (percent)	Share of boys (percent)	Difference (Boy – girl) (in percentage points)
Mathematics	45.42	33.62	34.71	53.84	19.13***
Physics	48.71	37.59	41.61	54.28	12.67***
Chemistry	48.91	37.61	41.81	54.49	12.68***
Biology	17.07	60.75	23.57	11.96	-11.61***
Computer Science	9.65	30.17	6.62	12.03	5.42***
History	15.58	54.89	19.43	12.55	-6.88***
Political Science	17.13	55.54	21.63	13.60	-8.03***
Geography	8.26	42.72	8.02	8.45	0.43***
Economics	35.18	47.57	38.04	32.93	-5.11***
Hindi	16.36	56.04	20.83	12.84	-7.99***
English	96.50	43.19	94.74	97.88	3.14***
Business Studies	29.57	46.03	30.94	28.50	-2.44***
Accounts	29.60	45.90	30.88	28.60	-2.28***
Subject combinations					
PCM (Phy+Chem+Math)	36.80	30.20	25.26	45.87	20.61***
Arts (History+Pol. Sc.)	13.00	54.09	15.98	10.66	-5.32***
Commerce (Bus. Studies+Acc)	29.17	46.03	30.52	28.11	-2.41***

Note: The table reports the pattern of stream choice of the most frequently offered subjects by CBSE schools post matriculation. The first column shows the subject/subject combination. The second column shows the percentage of all students who opt for that subject. The third column shows what percentage of students who opt for the subject are girls. The next two columns show the percentage of girls and boys, respectively, who opt for that subject. The last column reports differences between columns 5 and 4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 3.3: Gender wise class X scores

Subject	Overall mean (std. dev)	Girls' mean (std. dev)	Boys' mean (std. dev)	Difference (Boy – girl) (in std. dev terms)
English	70.51	71.40	69.82	-0.11***
	(14.83)	(15.08)	(14.59)	
Hindi	73.33	74.63	72.32	-0.17***
	(13.80)	(13.80)	(13.72)	
Mathematics	66.42	66.02	66.73	0.04***
	(17.45)	(17.35)	(17.52)	
Science	67.43	67.71	67.22	-0.03***
	(16.39)	(16.36)	(16.42)	
Social science	69.55	70.17	69.06	-0.07***
	(16.12)	(16.20)	(16.04)	
Total	348.72	351.45	346.58	-0.07***
	(73.96)	(74.25)	(73.66)	

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels in the last column indicate the statistical significance of the differences between columns 4 and 3. Standard deviations in parenthesis. Data source: Central Board of Secondary Education.

Table 3.4: Gender wise comparative advantage

	Girls	**Boys**	Difference **(Boy – girl)**
Mean standardized scores			
Mathematics	-0.023	0.018	0.04***
English	0.060	-0.047	-0.11***
Science	0.017	-0.013	-0.03***
Mean gap in scores			
Mathematics-English	-0.081	0.066	0.15***
Science-English	-0.041	0.035	0.08***
Mathematics-Science	-0.040	0.031	0.07***
Percent who score better in (%)			
Mathematics vs English	46.67	56.43	9.7***
Science vs English	49.42	55.19	5.77***
Mathematics vs Science	49.71	55.82	6.11***

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels in the last column indicate the statistical significance of the differences between columns 3 and 2. Scores are standardized by subtracting the cohort average from a student's own score and dividing the difference by the standard deviation of scores in the cohort. Data source: Central Board of Secondary Education.

Table 3.5: Gender composition and school type

Panel A

Type of school	**Number of schools**	**Number of students**	**Average female share in cohort (%) (std.dev)**
Total	12,685	24,03,957	43.99
			(23.75)
All girls	1046	2,17,994	1
			(0)
All boys	1345	1,64,083	0
			(0)
Co-education	10,474	20,21,880	41.53
			(13.01)

Panel B

Group	**Average female share in cohort (%) (std.dev)**
Girls	56.81
	(25.61)
Boys	33.92
	(16.16)
Girls in co-ed schools	45.60
	(14.71)
Boys in co-ed schools	38.63
	(10.75)

Note: Panel A gives summary statistics of different types of schools based on gender composition. Panel B reports gender wise peer composition of the cohort of an average student. Data source: Central Board of Secondary Education.

Table 3.6: Gender wise peer performance

Mean peer performance (out of 500)	Overall (std. dev)	For girls (std. dev)	For boys (std. dev)	Difference (Boy – girl)
Panel A: All schools				
All peers	348.71	344.94	351.67	6.74***
	(48.84)	(49.97)	(47.72)	
Female peers	359.79	351.41	367.28	15.87***
	(40.03)	(52.29)	(44.60)	
Male peers	347.03	347.80	346.55	-1.24***
	(48.40)	(48.76)	(48.16)	
Panel B: Co-educational schools				
All peers	356.69	354.16	358.48	4.33***
	(45.59)	(47.61)	(44.02)	
Female peers	365.22	362.32	367.28	4.96***
	(46.54)	(49.00)	(44.60)	
Male peers	350.64	347.80	352.66	4.86***
	(46.89)	(48.74)	(45.40)	

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels in the last column indicate the statistical significance of the differences between columns 4 and 3. Mean peer performance is calculated by subtracting a student's own score from the sum total of scores of the students in her cohort in her school, and dividing the difference by the number of students in her school cohort minus one. Data source: Central Board of Secondary Education.

Table 3.7: Stream choices of seniors

Role model measure	Overall mean (%) (std. dev)	Mean for girls (%) (std. dev)	Mean for boys (%) (std. dev)	Difference (Boy – girl)
Share of own gender in Math	51.96	41.09	59.58	18.48***
	(23.02)	(22.31)	(20.29)	
Share of own gender in PCM	43.26	31.43	51.54	20.12***
	(23.08)	(20.37)	(21.18)	
Share of own gender in Bio	18.81	27.93	12.43	-15.50***
	(17.66)	(19.53)	(12.81)	

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels in the last column indicate the statistical significance of the differences between columns 4 and 3. The role model variable is measured as the proportion of a student's own gender seniors who opted for the subject. Data source: Central Board of Secondary Education.

Table 3.8: Gender composition of senior class

Chilly climate measure	Overall mean (%) (std. dev)	Mean for girls (%) (std. dev)	Mean for boys (%) (std. dev)	Difference (Boy – girl)
Own gender in senior Math class	55.05	34.45	69.45	35.00***
	(23.20)	(16.81)	(14.58)	
Own gender in senior PCM class	55.43	31.59	72.06	40.47***
	(25.51)	(17.41)	(14.85)	
Own gender in senior Bio class	49.41	63.30	39.74	-23.56***
	(23.84)	(20.66)	(20.94)	

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels in the last column indicate the statistical significance of the differences between columns 4 and 3. The table shows the gender composition of students' seniors' classes. Data source: Central Board of Secondary Education.

Table 3.9: Gender wise socioeconomic characteristics

Attribute	Mean for girls (%)	Mean for boys (%)	Difference (Boy – girl)
Caste			
General	74.45	71.96	−2.49***
Scheduled Castes	7.70	7.00	−0.70***
Scheduled Tribes	3.60	3.10	−0.50***
Other Backward Classes	14.25	17.93	3.68***
Type of school administration			
Private aided	1.85	1.54	−0.31***
Government	24.83	15.40	−9.43***
Independent	62.37	72.49	10.12***
Jawahar Navodaya Vidyalaya	2.69	3.32	0.63***
Kendriya Vidyalaya	8.04	6.90	−1.14***
Type of board examination in class X			
External board exam	35.89	38.83	2.94***
School board exam	64.11	61.17	−2.94***
Other attributes			
Mean Annual family income (INR)	293,028.7	291,779	-1,249
(std. dev)	(6,96,698.5)	(22,48,034)	
Single child	5.23	6.21	0.98***

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Significance levels in the last column indicate the statistical significance of the differences between columns 3 and 2. Castes are the administrative caste categories of India. Data source: Central Board of Secondary Education.

Table 3.10: Socioeconomic attributes and stream choice

Attribute	Overall share (%)	Share of girls (%)	Share of boys (%)
Subject choice: Mathematics			
General	45.64	35.62	53.78
Scheduled castes	33.19	21.78	43.05
Scheduled Tribes	27.48	19.75	34.53
Other Backward Classes	53.60	40.71	61.65
External board exam	51.28	40.70	58.95
Single child	56.13	47.11	62.10
Subject choice: PCM			
General	35.54	24.64	44.40
Scheduled castes	27.82	16.46	37.62
Scheduled Tribes	25.42	17.87	32.30
Other Backward Classes	48.79	35.14	57.32
External board exam	44.40	32.92	52.74
Single child	47.66	37.26	54.55
Subject choice: Biology			
General	15.43	21.61	10.41
Scheduled castes	16.17	19.41	13.37
Scheduled Tribes	25.04	28.46	21.97
Other Backward Classes	23.17	34.78	15.92
External board exam	21.39	31.07	14.36
Single child	22.06	32.12	15.41

Note: Castes are the administrative caste categories of India. Data source: Central Board of Secondary Education.

Table 3.11: Stream choice and socioeconomic status

	(1) Math	(2) Math	(3) Math	(4) Pcm	(5) Pcm	(6) Pcm	(7) Bio	(8) Bio	(9) Bio
F	-0.191***	-0.183***	-0.176***	-0.206***	-0.197***	-0.191***	0.117***	0.121***	0.158***
	(0.00403)	(0.00356)	(0.00160)	(0.00384)	(0.00327)	(0.00152)	(0.00268)	(0.00229)	(0.00128)
SC		-0.110***	-0.0404***		-0.0603***	-0.0171***		0.0139***	0.0303***
		(0.00336)	(0.00157)		(0.00312)	(0.00147)		(0.00253)	(0.00120)
ST		-0.198***	-0.0792***		-0.125***	-0.0507***		0.0743***	0.0189***
		(0.00928)	(0.00371)		(0.00901)	(0.00382)		(0.00869)	(0.00309)
OBC		0.0628***	0.00960***		0.112***	0.0291***		0.0786***	0.0273***
		(0.00312)	(0.00123)		(0.00290)	(0.00121)		(0.00220)	(0.000958)
Annual family income		1.05e-08*	1.64e-09*		8.11e-09*	1.59e-09		2.63e-09*	8.25e-10*
		(5.55e-09)	(9.07e-10)		(4.46e-09)	(1.02e-09)		(1.47e-09)	(4.64e-10)
Single child		0.0951***	0.00156		0.0998***	0.0201***		0.0581***	0.0178***
		(0.00347)	(0.00151)		(0.00339)	(0.00154)		(0.00230)	(0.00122)
External board in X		0.0888***	0.0655***		0.112***	0.0617***		0.0652***	0.0209***
		(0.00415)	(0.00212)		(0.00391)	(0.00197)		(0.00251)	(0.00124)
Cohort FE	√	√	√	√	√	√	√	√	√
School FE			√			√			√
N	2384369	2384369	2384219	2384369	2384369	2384219	2384369	2384369	2384219
R^2	0.037	0.059	0.217	0.045	0.074	0.232	0.025	0.041	0.168

Note: Linear probability results are reported. Outcome is a dummy variable which takes value 1 if the subject is chosen. Robust standard errors clustered at the school level are in parenthesis. General category is the omitted caste category. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 3.12: Stream choice and ability

	(1) Math	(2) Math	(3) PCM	(4) PCM	(5) Bio	(6) Bio
Female	-0.176***	-0.191***	-0.191***	-0.205***	0.158***	0.135***
	(0.00160)	(0.00136)	(0.00152)	(0.00134)	(0.00128)	(0.00126)
Class X total score		-0.000435***		-0.000687***		0.00114***
		(0.0000386)		(0.0000373)		(0.0000295)
Class X math score		0.00424**		-0.00121		0.00375**
		(0.00206)		(0.00191)		(0.00158)
Class X science score		0.0142***		0.0189***		-0.00438***
		(0.00220)		(0.00203)		(0.00168)
CA in Math vs Eng		0.190***		0.212***		
		(0.0359)		(0.0333)		
CA in science vs Eng		-0.185***		-0.208***		-0.00178*
		(0.0359)		(0.0333)		(0.00108)
CA in science vs math						0.157***
						(0.0276)
SES	√	√	√	√	√	√
Cohort FE	√	√	√	√	√	√
School FE	√	√	√	√	√	√
N	2384219	2384219	2384219	2384219	2384219	2384219
R^2	0.217	0.417	0.232	0.389	0.168	0.202

Note: Linear probability results are reported. Outcome is a dummy variable which takes value 1 if the subject is chosen. Robust standard errors clustered at the school level are in parenthesis. CA stands for Comparative Advantage, measured as the difference in class X standardized scores of the two subjects. SES stands for socioeconomic status variables which include caste dummies, annual family income, single child status of the student and a dummy for taking external board exam in class X. General category is the omitted caste category. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 3.13: Gender gap decomposition: Ability

Blinder-Oaxaca decomposition						
Explanatory factor added	**Explained (percentage points) Percent explained (%)**					
	(1) Mathematics	(2) Mathematics	(3) PCM	(4) PCM	(5) Biology	(6) Biology
Total gap (percentage points)	19.08	19.08	20.57	20.57	-11.65	-11.65
SES	0.29***	0.27***	0.34***	0.33***	0.15***	0.14***
	(1.52%)	(1.42%)	(1.65%)	(1.60%)	(-1.29%)	(-1.20%)
Class X total score		0.22***		0.34***		-0.57***
		(1.15%)		(1.65%)		(4.89%)
Class X Math score		0.29		-0.08		0.26
		(1.52%)		(-0.39%)		(-2.23%)
Class X Science score		-0.73**		-0.97**		0.22
		(-3.83%)		(-4.72%)		(-1.89%)
CA in Math vs Eng		2.78***		3.10***		
		(14.57%)		(15.07%)		
CA in Science vs Eng		-1.40**		-1.57***		-0.01
		(-7.34%)		(-7.63%)		(0.09%)
CA in Science vs Math						-1.11***
						(9.53%)
Total	0.27***	1.42***	0.33***	1.14***	0.15***	-1.06***
	(1.42%)	(7.44%)	(1.60%)	(5.54%)	(-1.29%)	(9.10%)

Note: Decomposition results are reported. The first row shows the gender gap in take-up in percentage points for the column. The following rows report the detailed decomposition contribution of ability related variables in percentage points. The terms in the brackets report the contribution as a percent of the total gap shown in the top row. Robust standard errors clustered at the school level are in parenthesis. CA stands for Comparative Advantage, measured as the difference in class X standardized scores of the two subjects. SES stands for socioeconomic status variables which include caste dummies, annual family income, single child status of the student and a dummy for taking external board exam in class X. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 3.14: Stream choice and cohort peers

	(1) Math	(2) Math	(3) PCM	(4) PCM	(5) Bio	(6) Bio
Female	-0.176***	-0.176***	-0.191***	-0.191***	0.159***	0.158***
	(0.00160)	(0.00161)	(0.00152)	(0.00153)	(0.00128)	(0.00129)
Female share in class		0.0189**		0.0185**		0.0165**
		(0.00901)		(0.00870)		(0.00701)
Average peer score		-0.00000649		-0.0000116		-0.000253*
		(0.000210)		(0.000200)		(0.000149)
Average female peer score		0.000272***		0.000251***		0.000142**
		(0.0000914)		(0.0000845)		(0.0000653)
Average male peer score		0.000294**		0.000231*		0.000380***
		(0.000126)		(0.000122)		(0.0000903)
SES	√	√	√	√	√	√
Cohort FE	√	√	√	√	√	√
School FE	√	√	√	√	√	√
N	2007577	2006062	2007577	2006062	2007577	2006062
R^2	0.177	0.177	0.183	0.183	0.157	0.157

Note: Linear probability results are reported. Outcome is a dummy variable which takes value 1 if the subject is chosen. Robust standard errors clustered at the school level are in parenthesis. Sample is restricted to co-educational schools. SES stands for socioeconomic status variables which include caste dummies, annual family income, single child status of the student and a dummy for taking external board exam in class X. General category is the omitted caste category. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 3.15: Gender gap decomposition: Cohort peers

Blinder-Oaxaca decomposition			
Explanatory factor added	**Explained (percentage points) Percent explained (%)**		
	(1) Mathematics	(2) PCM	(3) Biology
Total gap (percentage points)	19.17	20.66	-14.90
Female share in cohort	-0.13	-0.13	-0.12
	(-0.68%)	(-0.63%)	(0.80%)
Average peer score	-0.003	-0.006	-0.11
	(-0.02%)	(-0.03%)	(0.74%)
Average female peer score	0.14	0.12	0.07
	(0.73%)	(0.58%)	(-0.47%)
Average male peer score	0.14	0.11	0.19
	(0.73%)	(0.53%)	(-1.28%)
SES	0.05	0.11**	0.06**
	(0.26%)	(0.53%)	(-0.40%)
Total	0.14	0.17	0.11
	(0.73%)	(0.82%)	(-0.74%)

Note: Decomposition results are reported. The first row shows the gender gap in take-up in percentage points for the column. The following rows report the detailed decomposition contribution of cohort-peer related variables in percentage points. The terms in the brackets report the contribution as a percent of the total gap shown in the top row. Robust standard errors clustered at the school level are in parenthesis. Sample is restricted to co-educational schools. SES stands for socioeconomic status variables which include caste dummies, annual family income, single child status of the student and a dummy for taking external board exam in class X. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 3.16: Stream choice and immediate seniors

	(1) Math	(2) Math	(3) PCM	(4) PCM	(5) (Bio)	(6) (Bio)
Female	-0.181***	-0.122***	-0.196***	-0.142***	0.162***	0.133***
	(0.00173)	(0.00206)	(0.00165)	(0.00212)	(0.00138)	(0.00170)
Share of own gender seniors in subject		0.0823***		0.111***		0.177***
		(0.00817)		(0.00864)		(0.00910)
Share of own gender in senior subject class		0.126***		0.0800***		0.00339
		(0.00520)		(0.00496)		(0.00311)
N	1309254	1309254	1301145	1301145	1289053	1289053
R^2	0.166	0.168	0.170	0.171	0.154	0.155
SES	√	√	√	√	√	√
Cohort FE	√	√	√	√	√	√
School FE	√	√	√	√	√	√

Note: Linear probability results are reported. Outcome is a dummy variable which takes value 1 if the subject is chosen. Robust standard errors clustered at the school level are in parenthesis. Sample is restricted to co-educational schools in the later two cohorts of 2015 and 2016. SES stands for socioeconomic status variables which include caste dummies, annual family income, single child status of the student and a dummy for taking external board exam in class X. General category is the omitted caste category. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 3.17: Gender gap decomposition: Immediate seniors

Blinder-Oaxaca decomposition			
Explanatory factor added	**Explained (percentage points) Percent explained (%)**		
	(1) Mathematics	(2) PCM	(3) Biology
Total gap (percentage points)	18.74	20.20	-15.87
Share of own gender seniors in subject	1.52	2.23***	-2.82*
	(8.11%)	(11.04%)	(17.77%)
Percent of own gender in senior subject class	4.41**	3.23***	-0.08
	(23.53%)	(15.99%)	(0.50%)
SES	-0.11*	-0.05	0.01
	(-0.59%)	(-0.25%)	(-0.06%)
Total	5.81***	5.41***	-2.89***
	(31.00%)	(26.78%)	(18.21%)

Note: Decomposition results are reported. The first row shows the gender gap in take-up in percentage points for the column. The following rows report the detailed decomposition contribution of variables related to immediate seniors in percentage points. The terms in the brackets report the contribution as a percent of the total gap shown in the top row. Robust standard errors clustered at the school level are in parenthesis. Sample is restricted to co-educational schools in the later two cohorts of 2015 and 2016. SES stands for socioeconomic status variables which include caste dummies, annual family income, single child status of the student and a dummy for taking external board exam in class X. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Chapter 4

Caste peer effects on student performance: Evidence from Indian schools

4.1 Introduction

Existence of peer effects in education is a common wisdom that is becoming an increasingly rigorously proven fact in the economics of education literature. A rich set of papers identifies the relationship between various peer attributes and student outcomes. These attributes can be peer ability and performance (Hanushek et al. 2003; Winston and Zimmerman 2004; Arcidiacono and Nicholson 2005; Griffith and Rask 2014; Antecol et al. 2016; Patacchini et al. 2017), peer background (McEwan 2003; Carrell and Hoekstra 2010; Rao 2019) or peer identity, like gender or race (Hoxby 2000; Angrist and Lang 2004; Hoxby and Weingarth 2005; Lavy and Schlosser

2011; Imberman et al. 2012; Billings et al. 2014; Oosterbeek and Van Ewijk 2014; Dewan et al. 2018)[1]. In this chapter, we look at the effect of the caste composition of a student's school peers on her academic performance in the context of India. Caste is the most important dimension of social division in India (Deshpande 2011; Munshi 2016, 2017), which makes our research question of particular interest and importance.

The institution of caste is a system of social stratification in the Indian society based on the ancient *Varna* system[2]. In this chapter castes are administrative categories based on the relative socioeconomic standings of various caste groups. Using three cohorts of student results data from the largest national education board in India, the Central Board of Secondary Education, we examine peer effects of students belonging to Scheduled castes (SC) and Scheduled tribes (ST), the most marginalized communities in the country, on test scores in national level standardized examination for students in the highest grade (class XII) in school. Using school fixed effects, cohort fixed effects and school-specific linear time trends, we rigorously show that the proportion of SC/STs in the peer group has an overall null effect on the test scores of students. Our results are precisely estimated, so that we can reject modest sized effects between 0.12σ and 0.14σ.

The issue of mixing individuals from different social and class backgrounds in an educational setting has always been a contentious one (Angrist and Lang 2004; Billings et al. 2014). In the Indian context, the Scheduled castes or the former "untouchables", situated at the lowest in the caste hierarchy, were historically deprived of any form of education (Deshpande 2011; Hnatkovska et al. 2012). After the af-

[1]See Sacerdote (2011) and Sacerdote (2014) for a excellent review of the literature on peer effects in education.

[2]See Deshpande (2011) for a comprehensive idea about the origin, evolution and contemporary relevance of the caste system in India.

firmative action policies were implemented post independence, the "untouchables" and the indigenous tribes of India were guaranteed quotas in political representation, government jobs and higher education[3]. More recently, the Right to Free and Compulsory Education Act (2009), commonly known as RTE, went one step further and mandated a minimum of 25% reserved seats for children of economically weaker and socially disadvantaged groups in all *primary* unaided private schools (Tucker and Sahgal 2012). This reopened the debate around reservations and its inefficiencies. Powerful private school lobbies filed a case in the Supreme Court of India, questioning the constitutional validity of the RTE act[4]. While the schools cited a potential loss in efficiency and autonomy due to government interference in the admission procedure (*mint* (New Delhi, 12 April, 2012)), surveys clearly brought out the apprehensions of parents and educators about the act (*Hindustan Times* (New Delhi, 19 September, 2012)). The survey showed that parents thought "the quality of education will go down as a result of the reservation" and that they were "unable to accept that my child and that my domestic help's child will be sitting next to each other in the same classroom". Though the RTE is not applicable in our setting because we look at test scores of students in the highest grade in school (and not the primary level where RTE is applicable), our motivation comes from the general perception and apprehension that the presence of disadvantaged caste students in a student's peer group will negatively affect her own performance (For recent evidence, see, for example, *News 18* (Saharanpur, 14 April, 2018); *The Print* (17 February, 2019)). In our sample of three cohorts of CBSE class XII students, we establish

[3]Affirmative action policies in India, commonly called the "reservation system" reserve 22.5 percent seats for the Scheduled Castes and the Scheduled Tribes in electoral constituencies, government jobs and institutes of higher education (Articles 15(4) and 16(4) of the Constitution of India). This percentage is roughly based on the share of SC/STs in the population of India.

[4]The act was finally upheld by the Supreme Court (Society for Unaided Private Schools of Rajasthan v Union of India & Another (2012) 6 SCC; Writ Petition (C) No.95 of 2010).

that a higher proportion of SC/ST students in the peer group has no negative (or positive) effect on the academic performance of individual students.

Identification of causal peer effects is tricky because students can select into schools endogenously. To address this, we use the methodology employed by Hoxby (2000) and identify the causal effects off the variation in the peer composition of adjacent cohorts within a school. By including school fixed effects in our empirical specification, we are able to eliminate the endogeneity bias stemming from self selection into schools in a given cohort. In addition, we include school-specific linear time trends to control for any time-varying unobservables at the school level. Finally, we also include a student's past scores as proxies for ability and past inputs into the education production function (Hanushek 1979).

Our results show that after controlling for a string of student socioeconomic characteristics, past scores, school and cohort fixed effects and school-specific linear time trends, there is no statistically significant effect of the cohort-to-cohort variation in the share of SC/STs in a student's peer group on her test scores in the national level standardized examination. This is in line with other studies which look at the peer effects of minority students (Angrist and Lang 2004; Hoxby and Weingarth 2005; Billings et al. 2014). We conduct a host of robustness checks and find that the null effects hold separately and are estimated precisely for all the caste groups, both genders, all income quartiles, both private and public schools and for students with different stream choices in higher secondary school. We also divide our sample into ability quartiles and check whether SC/ST peers of a certain ability have differential effects on students of a certain ability. We again find that the results are statistically insignificant. Most importantly, we show that lower ability SC/ST peers do not have any negative (or positive) impact on the test scores of students in any ability quar-

tile. These findings negate the most frequent apprehensions about SC/ST students lowering performance of the class by lowering the overall peer quality.

The literature on peer effects in education is vast and multidimensional. One dimension is the effects of attributes of peers like peer ability and peer performance. While Hanushek et al. (2003) and Winston and Zimmerman (2004) find that the relationship between peer performance and own performance is positive, Antecol et al. (2016) find adverse effects. Patacchini et al. (2017) also find positive effects of peer achievement, but show that the longevity of the effects is determined by the duration of friendship. Another dimension in the literature is the socioeconomic identity of peers. A multitude of studies look into the effects of female peers. Hoxby (2000), Lavy and Schlosser (2011) and Dewan et al. (2018) find positive gender peer effects on student performance, mediated mainly through an improvement in class discipline. Oosterbeek and Van Ewijk (2014), on the other hand, does not find substantial effects of female peers on student performance.

Our chapter is most closely related to studies which look at the race or ethnic dimensions of peer identity. Angrist and Lang (2004) look at the peer effects of black students in the context of the Metropolitan Council for Equal Opportunity (*Metco*, a race based desegregation program in the United States). They show that an increase in the proportion of *Metco* students, which are mostly blacks from Boston schools, has no impact on the scores of non-*Metco* students. Hoxby and Weingarth (2005) use natural experiments resulting from the reassignment of students in the Wake County school district, North Carolina. These reassignments were done with the goal of balancing schools' racial composition before the year 2000, and after that for balancing the schools' income composition. They show that after accounting for peer achievement, the race and income of peers have negligible effects on student

performance. Billings et al. (2014) utilize the end of race-based busing in Charlotte-Mecklenburg schools to study minority peer effects. Using the redrawing of school boundaries leading to new school assignments of students, they find that resegregation of schools due to end of race-based busing increased racial inequalities. Both white and minority students saw a worsening of outcomes when they were assigned to schools with higher proportion of minority students.

In the Indian context, the counterpart to race based peer effects in education is caste based peer effects in education. It is an emerging but still nascent field in India. Studies like Sekhri (2011, 2012) examine peer effects in the context of reservations in higher education; others like Sen et al. (2012) and Frijters et al. (2017, 2019) use exogenous variation in peer assignment to study the effect of the caste identity of peers in tertiary education. In a recent paper, García-Brazales et al. (2020) look at the effect of the caste composition of classroom on middle-school students in the Indian state of Andhra Pradesh. All these studies, however, are limited in geographic scope. To the best of our knowledge, this study is the first to look at caste peer effects at the school level using three cohorts of rich administrative data from the largest education board in the country with an all-India presence.

The rest of the chapter is organised as follows: Section 4.2 gives the institutional background about the schooling system and the caste system in India, followed by a description on the data in Section 4.3. Section 4.4 elaborates on our Empirical Strategy. Section 4.5 presents our results and robustness checks, Section 4.6 concludes.

4.2 Institutional Background

The school system of India follows a 10+2+3 structure, as described in section 3.2. This includes 10 years of schooling up to the secondary level (class X), followed by 2

years of higher secondary schooling (class XI and XII) and three years of graduation. The class XII board examination is a high stake test as the scores are crucial for admission into courses for higher education into various colleges and universities. They are arguably the most important examinations a student appears for in her school life.

In this chapter, we look at effect of the proportion of Scheduled castes (SC) and Scheduled tribes (ST) in the peer group on a students's performance[5]. The SC and ST are the most socioeconomically marginalized communities in the country. For example, in the 68^{th} round of the the National Sample Survey (NSS) conducted in 2011-12, the overall literacy rate in the country was 74.1%. The rate among the General category was 83.2% and among the OBCs was 72.8%. In contrast, only 67.4% SCs and 64.8% STs were literate in 2011-12 (NSS Report, 2015).

4.3 Data

We use the three cohorts of board examination results data from the CBSE, the largest national education board in India[6]. In this chapter, we use a student's covariates at the class XII level. We use the class X roll number of class XII students to map them to their class X scores, which we use as proxies for a student's ability and past inputs. The data also provides information on the socioeconomic characteristics of the students. The caste status is used to construct our variable of interest.

Table 4.1 provides summary statistics for our sample of students which includes only school-cohorts which have a strictly positive but below 100% share of SC/ST students[7]. Out of a total of 2,152,475 students in our sample, 8.82% are SC and

[5]Refer to the Appendix to Chapter 2 for a brief overview of the caste system in India.
[6]Refer to section 3.3 for details on the CBSE data.
[7]This restriction drops 208,690 or 8.84% of the total observations.

3.56% are ST. OBCs form 18.50% of the sample. The rest 69.12% are formed by the residual category, often termed as the "General" category. The mean score out of a total of 500 is 328.68 in class XII board examination and 347.53 in class X board examination. Our main variable of interest, the proportion of SC/ST in the school-cohort, has a mean value of 0.1262 or 12.62%.

4.4 Empirical Strategy

Our aim in this chapter is to explore caste peer effects on student performance in class XII board examinations. We use three cohorts of students who appeared in class XII board examinations under the CBSE in 2014, 2015 and 2016. Since the peer composition across schools in a cohort is potentially endogenous, we exploit cohort to cohort variation in peer composition of a student in a given grade within a school to elicit the effect. While a student or her family may choose a school depending upon its locality and the expected proportion of students from various castes, the exact realisations of these variables in the particular year the student takes admission can not be predicted. As pointed out by Hoxby (2000), idiosyncratic variations in the peer compositions of adjacent cohorts in a given school could be the result of demographic variations in a school's catchment area and are arguably exogenous.

We use the following regression specification:

$$S_{isc} = \beta_0 + \beta_1.Pscst_{isc} + \beta_2.X_{isc} + C_c + sch_s + L_s.c + \varepsilon_{isc}. \tag{4.1}$$

Here our dependent variable, S_{isc}, is the standardized value of the total score in class XII board examination of student i in school s in cohort c. Scores are standardized by subtracting the cohort mean of scores from a student's own score

and dividing the difference by the cohort standard deviation. Our variable of interest, $Pscst_{isc}$, is the proportion of students belonging to Scheduled castes or Scheduled tribes in the peers of student i in school s in cohort c[8]. The third term, X_{isc} is a vector of student level characteristics including her gender, caste status, single child status, annual family income, and her choice of board examination in class X (board based or school based). Next we include cohort fixed effects, C_c, to control for any unobservables within a cohort. School fixed effects, sch_s, are included to remove any bias stemming from endogenous selection of schools in a cohort. There may still be time-varying unobservables at the school level which are correlated with the proportion of SC/ST students in the peer group and the student's test scores. For this, we include school specific linear time trends, $L_s.t$, as another set of controls. Thus, our identification comes from the deviation in the proportion of SC/ST students in a school-cohort from the school specific linear time trend. Our estimates will be causal as long as the time-varying unobservables at the school level follow a linear trend.

To further strengthen our claim to causality, we also include a student's past score in class X board examination as a proxy for her ability as well as a measure of past inputs into the student's education production function[9]. Students can sort into schools according to ability. Failure to include a measure of ability could lead to omitted variable bias since ability is likely to be correlated with both peer composition (via selection of schools) and performance of a student. In addition, the observed performance of a student at any point in time is an outcome of inputs

[8]It is important to note that this proportion is calculated based on the entire cohort of a student in her school, irrespective of whether the peers took their class X board examination under the CBSE or under some other board.

[9]It is here that our sample gets restricted to students who appeared for both class X and XII board examination under the CBSE.

put in by the student herself, her family, peers, instructors and school, among other things (Hanushek 1979). This will again lead to biased estimates since these inputs will be correlated with peer composition as well as the outcome variable. The score of students in class X can serve as a plausible proxy for both these factors.

4.5 Results

4.5.1 Effect of SC/ST proportion on student scores

Table 4.2 presents our main results. The dependent variable in each column is the standardized total score of class XII board examination. The main variable of interest is the proportion of students belonging to Scheduled castes or Scheduled tribes in a student's cohort in her school. The first column in Table 4.2 runs the OLS model with only cohort fixed effects. We see that the coefficient on the proportion of SC/ST students in the cohort is negative and statistically significant. However, as discussed in the previous section, selection into schools *within* a cohort is possibly endogenous. To eliminate this bias, school fixed effects are included in the next column. The coefficient on our variable of interest increases in magnitude to 0.117, but is now statistically significant at only 5%.

Finally, in column 3, we include school-wise linear time trends in our specification. This guards against time-varying unobservables at the school level to the extent that they follow a linear trend. The addition of the linear trends not only makes the coefficient on the proportion of SC/ST statistically insignificant, but it also reduces the size of the coefficient by a factor of 100. The point estimate has high precision and we can reject effects ranging between 0.12σ and 0.14σ.

To put our caste peer effect coefficient into perspective, we compare them to the

gradients of other covariates. The coefficients on the other controls shown in the table are larger in size and always statistically significant. For example, the coefficient on class X score of a student hovers around 0.6 and the peer effects coefficient is only 1.3% of this coefficient. Thus, we can confidently rule out meaningful effects of SC/ST peers on test scores.

The coefficient on the female student dummy is always positive as girls have a higher score compared to boys on an average. The caste dummies show that the scores of students of all other caste categories are lower than the omitted General category. Other controls not shown in the table include a student's annual family income, her only child status and a dummy for whether she appeared for a school based or a board based examination in class X.

Thus, after controlling for a string of observables at the student level (gender, caste status, annual family income, single child status and choice of board examination), school and cohort fixed effects, as well as school-specific linear time trends, we find that the proportion of SC/STs in the peer group has no statistically significant effect on student performance in class XII board examination in our sample.

Next we explore whether the proportion of SC/STs in the peer group has any heterogeneous effects in different sub-samples. Table 4.3 presents the results by caste groups and Table 4.4 presents them by gender groups. The columns in Table 4.3 look at the effect of proportion of SC/STs in peer group for students belonging to SC/ST caste group, to OBC caste group and to the residual General category caste group. As can be seen, the coefficients are statistically insignificant for each of these groups. The columns in Table 4.4 run our preferred specification on the sample of girls and boys separately. Here again we see that the proportion of SC/STs in the peer group has no statistically significant effect on either of the genders.

In Table 4.5 we report results by income quartiles. We create annual family income quartiles at the district level within a cohort and run our preferred specification on each quartile separately. Again, we see that although the magnitude of the coefficient on the proportion of SC/STs in the peer group varies, it is not statistically significant for any income quartile. Thus, the null effect of proportion of SC/ST students in the peer group is consistent across all the above sub-samples.

In the next table, results are reported separately for private and public schools. Our private school definition includes aided and unaided independent schools. The public schools include government schools as well as Kendriya Vidyalays and Jawahar Navodaya Vidyalayas, both of which are also run by a government administration. Table 4.6 shows that our null results hold for both private and public schools. The coefficient on the proportion of SC/ST students in the peer group is statistically insignificant in each column. Thus, the proportion of SC/ST students in the peer group does not have any statistically significant effect on student performance in the overall sample, for any of the caste groups, for any of the genders, for any of the income quartiles and for any type of the schools.

4.5.2 Non linear effects by ability

In this subsection, we examine whether our null results mask heterogeneity by ability. It could be possible, for example, that while the proportion of SC/ST students does not have any effect on student performance on an average, the effect may differ by the ability of those students. If the direction of this effect is opposite for lower ability students as compared to higher ability students, we might see a null effect on average. We divide the students into quartiles by cohort according to their class X scores and run our preferred specification for each quartile. Table 4.7 reports the results. The

first column has the sample of students in the lowest ability quartile, the last column has those in the highest ability quartile. We see that the coefficient is statistically insignificant for all quartiles: the effect of the proportion of SC/ST students in the peer group does not differ by student ability and is null for all ability quartiles.

We then take a step further and attempt to elicit non-linear effects of the proportion of SC/ST students in peer group. This directly addresses the apprehensions discussed in the introduction about SC/ST peers having negative externalities due to their lower academic performance on average. In particular, we want to answer questions like "Does the proportion of low ability SC/ST students in the peer group have a negative effect on the performance of students who themselves are of lower ability?" or "Do high achieving SC/ST peers have a differential effect on the performance of high ability students?". Following Imberman et al. (2012), we assign an ability quartile to each student as in Table 4.7. Then, for each ability quartile, we calculate the proportion of SC/ST peers. Thus, for each student, we assign her ability quartile, the proportion of SC/ST peers in her own ability quartile, as well as the proportion of SC/ST peers in the other three ability quartiles. We then run our preferred specification for each of these quartiles.

The results are presented in Table 4.8. Column 1 has the full sample of students, column 2 has the students in the lowest ability quartile, followed by the next quartiles in columns 3 and 4 and the highest quartile in column 5. Our independent variable is also divided into four independent variables as described above. In column 1, we see that the coefficients on the proportion of SC/STs in the peer group are statistically insignificant irrespective of the ability quartile. More importantly, rows 1 and 2 reveal that the lowest ability SC/ST peers have no effect on the performance of students in any of the ability quartiles. In fact, the coefficients on our variables of

interest are statistically insignificant in 17 out of the 20 cells. The three instances of a negative effect are both small in size and low in statistical significance.

Summing up, we see that the proportion of SC/ST students in the peer group has a consistently null effect on student performance in our sample. This subsection shows that the null effects seen in the previous subsection do not mask heterogeneous effects by either student ability or peer ability- even low-ability SC/ST peers do not impose a negative peer effect.

4.5.3 Robustness checks

In this subsection, we present a battery of robustness checks to further strengthen our results. Table 4.9 presents the first set of these results. In column 1, we add a quadratic term of the proportion of SC/ST students in the peer group to detect any non-linear effects of the variable. We see that the coefficient of the quadratic term is statistically insignificant. In addition, it has little effect on the magnitude or the statistical significance of our main variable of interest.

In the second column, we add, as additional controls, the mean past scores (class X board examination scores) of all the SC/ST peers of a student as well as the mean past scores of all peers of a student. The coefficient on the proportion of SC/ST now turns negative but it is still statistically insignificant[10].

In Table 4.10, we check if our results hold for students who self select into various subject streams. After matriculation in class X, students in India are required to choose specialized subjects for the next two years of study (classes XI and XII). We group students into the most common subject combinations opted by them in

[10]In our dataset, the past scores are available only for students who appeared for their class X board examinations also under the CBSE, and not for the entire cohort of a student which includes students who migrated from other education boards.

our data. The Science stream comprises of students who opt for both Physics and Chemistry, Commerce stream comprises of students who opt for Business Studies and Accounts, and Arts stream has students who choose History and Political Science, among other subjects. In general, a student does not choose subjects from any two of the above streams simultaneously; thus they provide meaningful slices of the data. We see from Table 4.10 that the proportion of SC/ST students in the peer group has a null effect for each of the three stream choices.

Finally, Table 4.11 presents results for student groups depending upon whether they changed schools after class X. About 30% of students in our data change schools between class X and XII, most likely because many schools have grades only till class X. While a change in schools after class X could be a necessity, it can confound our estimates if the school changes are made in response to the peer composition of a school. Thus we divide students into those who did change schools between class X and XII and those who did not, and see if our results differ across the two groups. We again see that the coefficient on the proportion of SC/STs in the peer group is statistically insignificant for both categories of students.

The robustness checks presented in this subsection rigorously demonstrate that the proportion of SC/STs in the cohort peers has consistent and precisely estimated null effect on the performance of students in class XII.

4.6 Conclusion

We estimate caste peer effects of students from the most disadvantaged castes on test scores of class XII students in national level standardized examination using three cohorts of results data from the largest education board in the country, the Central Board of Secondary Education. We use cohort-to-cohort variation in the share of

SC/ST students in the peer group of students within a school to causally identify its effects on student performance. By controlling for socioeconomic characteristics of students, their past scores, cohort and school fixed effects, and school-specific linear time trends, we find robust evidence of an overall null effect of the proportion of SC/ST students in a student's school peers on her own academic score. Our point estimates are precise and we can reject modest sized estimates between 0.12σ and 0.14σ. Our results hold separately for all caste categories, both genders, all four income quartiles and for all students in various streams in class XII. It also holds for students who changed schools after class X and for those who did not. Lastly, we find evidence of small negative effects of the highest ability SC/ST peers on high ability students. It is possible that a student's narrow peer group is defined within her school with select students of her class. We present our results with this caveat in mind that what we call peers may be too broad a definition. However, our definition of peers at the school-cohort level is in line with the motivation with which we started. The common apprehensions about SC/ST students start with them sitting in the same classrooms as higher caste students, regardless of whether they actually mingle with those students. Our results establish that the most cited reason for this apprehension- negative effect on student performance- is not true in our sample.

We believe there are interesting policy and social implications of our findings. Although we do not study the effects of policies which aim at redistribution of students across schools per se, our results suggest that an increase in students from marginalized castes in a cohort have no detrimental effect on student performance. Our results send a strong social message because they show evidence against the popular myth that the disadvantaged castes necessarily have negative externalities

due to their lower academic performance on average.

Tables for Chapter 4

Table 4.1: Summary Statistics

Total students	2,152,475
Girls	44.13%
Caste	
Scheduled Castes	8.82%
Scheduled Tribes	3.56%
Other Backward Castes	18.50%
General	69.12%
Other attributes	
External board exam	36.56%
Mean Annual family income (INR)	270,372.6
(std. dev)	(898,262.9)
Single child	5.31%
Mean score (out of 500) **(std. dev)**	
Class XII	328.68
	(82.62)
Class X	347.53
	(73.84)
Mean proportion of SC/ST in peer group	0.1262
(std.dev)	(0.1507)

Note: Castes are the administrative caste categories in India. Data source: Central Board of Secondary Education.

Table 4.2: Caste peer effects and class XII score

	(1)	(2)	(3)
Proportion of SC/ST students	-0.0994***	-0.117**	0.00796
	(0.0362)	(0.0558)	(0.0669)
Class X total score	0.643***	0.607***	0.614***
	(0.00422)	(0.00257)	(0.00259)
Female	0.220***	0.138***	0.135***
	(0.00608)	(0.00229)	(0.00229)
SC	-0.0742***	-0.0597***	-0.0584***
	(0.00444)	(0.00198)	(0.00197)
OBC	-0.156***	-0.0444***	-0.0447***
	(0.00777)	(0.00198)	(0.00192)
ST	-0.113***	-0.0767***	-0.0750***
	(0.0145)	(0.00451)	(0.00445)
Cohort FE	√	√	√
School FE		√	√
School-wise time trend			√
N	2077833	2077810	2077810
R^2	0.452	0.614	0.628

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.3: Heterogeneity by caste

	(1) SC/ST	(2) OBC	(3) General
Proportion of SC/ST students	-0.0633	-0.0473	0.0288
	(0.0943)	(0.109)	(0.0766)
Class X total score	0.545***	0.629***	0.621***
	(0.00467)	(0.00359)	(0.00277)
Female	0.129***	0.161***	0.128***
	(0.00451)	(0.00372)	(0.00243)
Cohort FE	√	√	√
School FE	√	√	√
School-wise linear trend	√	√	√
N	256564	382780	1437301
R^2	0.674	0.622	0.632

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. The sample for each column is specified at the top. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.4: Heterogeneity by gender

	(1) Girls	(2) Boys
Proportion of SC/ST students	0.00300	0.0149
	(0.0824)	(0.0809)
Class X total score	0.585***	0.633***
	(0.00300)	(0.00295)
SC	-0.0613***	-0.0575***
	(0.00254)	(0.00278)
OBC	-0.0326***	-0.0505***
	(0.00236)	(0.00235)
ST	-0.0626***	-0.0858***
	(0.00552)	(0.00559)
Cohort FE	√	√
School FE	√	√
School-wise linear trend	√	√
N	921765	1155931
R^2	0.637	0.633

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. The sample for each column is specified at the top. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.5: Heterogeneity by income quartiles

	(1) Quartile 1	(2) Quartile 2	(3) Quartile 3	(4) Quartile 4
Proportion of SC/ST students	0.112	0.0190	0.0222	-0.143
	(0.0978)	(0.0975)	(0.0993)	(0.103)
Class X total score	0.630***	0.618***	0.610***	0.603***
	(0.00356)	(0.00340)	(0.00316)	(0.00360)
Female	0.162***	0.154***	0.129***	0.0980***
	(0.00354)	(0.00331)	(0.00313)	(0.00301)
SC	-0.0431***	-0.0549***	-0.0685***	-0.0824***
	(0.00302)	(0.00358)	(0.00381)	(0.00430)
OBC	-0.0337***	-0.0404***	-0.0456***	-0.0460***
	(0.00298)	(0.00327)	(0.00302)	(0.00326)
ST	-0.0560***	-0.0775***	-0.0710***	-0.0955***
	(0.00714)	(0.00820)	(0.00727)	(0.00681)
Cohort FE	√	√	√	√
School FE	√	√	√	√
School-wise linear trend	√	√	√	√
N	593438	518597	494177	470203
R^2	0.624	0.614	0.628	0.693

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. Quartile 1 corresponds to the poorest income quartile, quartile 4 corresponds to the richest. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.6: Heterogeneity by school administration

	(1) Private	(2) Public
Proportion of SC/ST students	0.00271	-0.000445
	(0.108)	(0.0829)
Class X total score	0.644***	0.530***
	(0.00300)	(0.00443)
Female	0.145***	0.0861***
	(0.00274)	(0.00352)
SC	-0.0851***	-0.0465***
	(0.00305)	(0.00250)
OBC	-0.0564***	0.000506
	(0.00233)	(0.00270)
ST	-0.119***	-0.0392***
	(0.00651)	(0.00554)
Cohort FE	√	√
School FE	√	√
School-wise linear trend	√	√
N	1381105	695550
R^2	0.630	0.628

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. The sample for each column is specified at the top. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.7: Heterogeneity by quartiles of student ability

	(1) Quartile 1	(2) Quartile 2	(3) Quartile 3	(4) Quartile 4
Proportion of SC/ST students	-0.130	-0.0412	0.0923	-0.0301
	(0.114)	(0.0966)	(0.0895)	(0.0875)
Class X total score	0.654***	0.562***	0.600***	0.939***
	(0.00735)	(0.00586)	(0.00504)	(0.00640)
Female	0.150***	0.194***	0.147***	0.0301***
	(0.00400)	(0.00328)	(0.00305)	(0.00275)
Cohort FE	√	√	√	√
School FE	√	√	√	√
School-wise linear trend	√	√	√	√
N	535134	523533	513771	504159
R^2	0.727	0.485	0.461	0.791

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. Quartile 1 corresponds to students with the lowest class X scores, quartile 4 corresponds to those with the highest class X scores. Other controls in all columns include caste dummies, annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.8: Non linear effects by ability quartiles

	(1) Full sample	(2) Quartile 1	(3) Quartile 2	(4) Quartile 3	(5) Quartile 4
Proportion of SC/ST students in Q1	0.0145	0.0544	0.0210	0.0263	-0.0193
	(0.0234)	(0.0523)	(0.0340)	(0.0294)	(0.0312)
Proportion of SC/ST students in Q2	-0.00144	-0.0579	-0.0781	0.0726	0.0156
	(0.0397)	(0.0776)	(0.0586)	(0.0516)	(0.0530)
Proportion of SC/ST students in Q3	-0.0240	0.0294	-0.103*	-0.0676	-0.00479
	(0.0421)	(0.0729)	(0.0559)	(0.0543)	(0.0577)
Proportion of SC/ST students in Q4	-0.0147	0.0128	0.0155	-0.0822*	-0.105*
	(0.0345)	(0.0446)	(0.0587)	(0.0421)	(0.0567)
Class X total score	0.612***	0.652***	0.553***	0.602***	0.931***
	(0.00272)	(0.00800)	(0.00598)	(0.00534)	(0.00695)
Cohort FE	√	√	√	√	√
School FE	√	√	√	√	√
School-wise time trend	√	√	√	√	√
N	1742711	396620	455035	461668	428430
R^2	0.609	0.715	0.477	0.454	0.780

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. Quartile 1 corresponds to students with the lowest class X scores, quartile 4 corresponds to those with the highest class X scores. The independent variables are the proportion of SC/ST students in each quartile of student performance. Other controls in all columns include a female student dummy, caste dummies, annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.9: Robustness I: Inclusion of quadratic term and mean ability controls

	(1)	(2)
Proportion of SC/ST students	0.00741	-0.0811
	(0.117)	(0.0628)
(Proportion of SC/ST students) sq	0.00111	
	(0.197)	
Class X total score	0.614***	0.614***
	(0.00259)	(0.00261)
Female	0.135***	0.135***
	(0.00229)	(0.00231)
SC	-0.0584***	-0.0568***
	(0.00197)	(0.00198)
OBC	-0.0447***	-0.0441***
	(0.00192)	(0.00189)
ST	-0.0750***	-0.0734***
	(0.00444)	(0.00445)
Mean controls		√
Cohort FE	√	√
School FE	√	√
School-wise time trend	√	√
N	2077810	2047950
R^2	0.628	0.628

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. Mean controls include mean past scores of SC/ST peers and mean past scores of all peers in the cohort. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.10: Robustness II: Different stream choices of students

	(1) Science	(2) Commerce	(3) Arts
Proportion of SC/ST students	0.0375	-0.103	-0.0811
	(0.0811)	(0.110)	(0.166)
Class X total score	0.746***	0.728***	0.754***
	(0.00306)	(0.00320)	(0.00884)
Female	0.0672***	0.102***	0.103***
	(0.00282)	(0.00336)	(0.00672)
SC	-0.0465***	-0.0853***	-0.0359***
	(0.00272)	(0.00393)	(0.00352)
OBC	-0.0276***	-0.0512***	0.0242***
	(0.00236)	(0.00308)	(0.00534)
ST	-0.0799***	-0.144***	0.0142
	(0.00473)	(0.00769)	(0.0107)
Cohort FE	√	√	√
School FE	√	√	√
School-wise linear trend	√	√	√
N	994354	591334	287246
R^2	0.668	0.643	0.710

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. The three columns report results for students in different streams in class XII. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Table 4.11: Robustness III: School changing status of students

	(1) Changed schools after class X	(2) Did not change schools after class X
Proportion of SC/ST students	-0.0694	0.0268
	(0.113)	(0.0721)
Class X total score	0.579***	0.666***
	(0.00423)	(0.00288)
Female	0.165***	0.112***
	(0.00410)	(0.00213)
SC	-0.0815***	-0.0445***
	(0.00402)	(0.00202)
OBC	-0.0496***	-0.0315***
	(0.00337)	(0.00191)
ST	-0.0982***	-0.0487***
	(0.00766)	(0.00464)
Cohort FE	√	√
School FE	√	√
School-wise linear trend	√	√
N	624080	1453469
R^2	0.600	0.668

Note: Linear regression results are reported. Outcome is standardized value of student's class XII total score. The sample in column 1 consists of students who changed their schools after class X, and the sample in the second column has students who remained in the same school as they were in class X. Other controls in all columns include annual family income, only child status and a dummy for whether the student appeared for a board-based or a school-based class X exam. Robust standard errors clustered at the school level are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Data source: Central Board of Secondary Education.

Bibliography

Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of Research Design*, 1(8):1–8.

Aghion, P., Boustan, L., Hoxby, C., and Vandenbussche, J. (2009). The causal impact of education on economic growth: Evidence from US. *Brookings papers on economic activity*, 1:1–73.

Ahuja, A. and Ostermann, S. L. (2016). Crossing caste boundaries in the modern Indian marriage market. *Studies in Comparative International Development*, 51(3):365–387.

Ajayi, K. F. and Buessing, M. (2015). Gender parity and schooling choices. *The Journal of Development Studies*, 51(5):503–522.

Akbari, M., Bahrami-Rad, D., and Kimbrough, E. O. (2019). Kinship, fractionalization and corruption. *Journal of Economic Behavior & Organization*, 166:493–528.

Akbari, M., Bahrami-Rad, D., Kimbrough, E. O., Romero, P. P., and Alhosseini, S. (2020). An experimental study of kin and ethnic favoritism. *Economic Inquiry*, 58(4):1795–1812.

Allendorf, K. and Pandian, R. K. (2016). The decline of arranged marriage? Marital

change and continuity in India. *Population and Development Review*, 42(3):435–464.

Allendorf, K. and Thornton, A. (2015). Caste and Choice: The influence of Developmental Idealism on Marriage Behavior. *American Journal of Sociology*, 121(1):243–287.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *Journal of political economy*, 113(1):151–184.

Ambedkar, B. R. (1936). *Annihilation of Caste*. Bheema Patrika Publications, Jallandhar, India.

Ambedkar Scheme for Social Integration through Inter-Caste Marriages (2016). *Dr. Ambedkar Foundation Ministry of Social Justice and Empowerment*.

Anderson, J. and Lightfoot, A. (2019). The school education system in India: An Overview. *The British Council*.

Anderson, S. (2011). Caste as an Impediment to Trade. *American Economic Journal: Applied Economics*, 3(1):239–63.

Anelli, M. and Peri, G. (2019). The effects of high school peers' gender on college major, college performance and income. *The Economic Journal*, 129(618):553–602.

Angrist, J. D. and Lang, K. (2004). Does school integration generate peer effects? Evidence from Boston's Metco Program. *American Economic Review*, 94(5):1613–1634.

Antecol, H., Eren, O., and Ozbeklik, S. (2016). Peer effects in disadvantaged primary schools: Evidence from a randomized experiment. *Journal of Human Resources*, 51(1):95–132.

Arcidiacono, P. and Nicholson, S. (2005). Peer effects in medical school. *Journal of Public Economics*, 89(2-3):327–350.

Ashraf, J. and Ashraf, B. (1993). Estimating the gender wage gap in Rawalpindi city. *The Journal of Development Studies*, 29(2):365–376.

Astorne-Figari, C. and Speer, J. D. (2019). Are changes of major major changes? The roles of grades, gender, and preferences in college major switching. *Economics of Education Review*, 70:75–93.

Banerjee, A., Duflo, E., Ghatak, M., and Lafortune, J. (2013). Marry for what? Caste and mate selection in modern India. *American Economic Journal: Microeconomics*, 5(2):33–72.

Banerji, M. (2008). Is education associated with a transition towards autonomy in partner choice? A case study of India. Master's thesis, University of Maryland.

Banerji, M., Martin, S., and Desai, S. (2013). Are the young and educated more likely to have 'love' than arranged marriage? A study of autonomy of partner choice in India. *Working Paper Series (pp. 1À43), NCAER, New Delhi.*

Baram-Tsabari, A. and Yarden, A. (2011). Quantifying the gender gap in science interests. *International Journal of Science and Mathematics Education*, 9(3):523–550.

Barro, R. J. (2001). Education and economic growth. In *The contribution of human*

and social capital to sustained economic growth and well-being, pages 14–41. OECD and Human Resources Development, Canada.

Basu, S. (2015). Intermarriage and the labor market outcomes of Asian women. *Economic Inquiry*, 53(4):1718–1734.

Beegle, K., Frankenberg, E., and Thomas, D. (2001). Bargaining power within couples and use of prenatal and delivery care in Indonesia. *Studies in Family Planning*, 32(2):130–146.

Belfield, C., Britton, J., Buscha, F., Dearden, L., Dickson, M., Van Der Erve, L., Sibieta, L., Vignoles, A., Walker, I., and Zhu, Y. (2018). The relative labour market returns to different degrees: Research report. *Institute for Fiscal Studies*.

Beteille, A. (1971). Race, caste and ethnic identity. *International Social Science Journal*, 23(4):519–535.

Bettinger, E. P. and 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.

Bidner, C. and Eswaran, M. (2015). A gender-based theory of the origin of the caste system of India. *Journal of Development Economics*, 114:142–158.

Billings, S. B., Deming, D. J., and Rockoff, J. (2014). School segregation, educational attainment, and crime: Evidence from the end of busing in Charlotte-Mecklenburg. *The Quarterly Journal of Economics*, 129(1):435–476.

Bisin, A. and Verdier, T. (2000). "Beyond the melting pot": Cultural transmission, marriage, and the evolution of ethnic and religious traits. *The Quarterly Journal of Economics*, 115(3):955–988.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, Oct(1):436–455.

Boll, C. and Lagemann, A. (2018). Gender pay gap in EU countries based on SES (2014). *Directorate-General for Justice, Luxembourg, Publication Office of the European Union.*, doi:10.2838/978935.

Bostwick, V. K. and Weinberg, B. A. (2018). Nevertheless she persisted? Gender peer effects in doctoral STEM programs. *NBER Working Paper No. w25028*.

Bottia, M. C., Stearns, E., Mickelson, R. A., Moller, S., and Valentino, L. (2015). Growing the roots of STEM majors: Female Math and Science high school faculty and the participation of students in STEM. *Economics of Education Review*, 45:14–27.

Bronson, M. A. (2014). Degrees are forever: Marriage, educational investment, and lifecycle labor decisions of men and women. *Unpublished manuscript (2), Department of Economics, UCLA*.

Brown, C. J., Pagán, J. A., and Rodríguez-Oreggia, E. (1999). Occupational attainment and gender earnings differentials in Mexico. *ILR Review*, 53(1):123–135.

Buonanno, P. and Pozzoli, D. (2009). Early labour market returns to college subject. *Labour*, 23(4):559–588.

Carl, N. (2017). Ethnicity and electoral fraud in Britain. *Electoral Studies*, 50:128–136.

Carrell, S. E. and Hoekstra, M. L. (2010). Externalities in the classroom: How children exposed to domestic violence affect everyone's kids. *American Economic Journal: Applied Economics*, 2(1):211–28.

Chakrabarti, A. (2009). Determinants of participation in higher education and choice of disciplines: Evidence from urban and rural Indian youth. *South Asia Economic Journal*, 10(2):371–402.

Chanana, K. (2000). Treading the hallowed halls: Women in higher education in India. *Economic and Political Weekly*, pages 1012–1022.

Chanana, K. (2007). Globalisation, higher education and gender: Changing subject choices of Indian women students. *Economic and Political Weekly*, pages 590–598.

Charness, G. and Gneezy, U. (2012). Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58.

Cheney, G. R., Ruzzi, B. B., and Muralidharan, K. (2005). A profile of the Indian education system. *Prepared for the New Commission on the Skills of the American Workforce, National Center on Education and the Economy*.

Chi, W. and Li, B. (2014). Trends in China's gender employment and pay gap: Estimating gender pay gaps with employment selection. *Journal of Comparative Economics*, 42(3):708–725.

Chowdhry, P. (1997). Enforcing cultural codes: Gender and violence in northern India. *Economic and Political Weekly*, 32(19):1019–1028.

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

Cutler, D. M. and Lleras-Muney, A. (2006). Education and health: Evaluating theories and evidence. *NBER Working No. w12352*.

Dahl, G. B., Rooth, D.-O., and Stenberg, A. (2020). Long-run returns to field of study in secondary school. *NBER Working Paper No. w27524*.

Das, P. (2012). Wage inequality in India: Decomposition by sector, gender and activity status. *Economic and Political Weekly*, pages 58–64.

Davis, K. (1941). Intermarriage in caste societies. *American Anthropologist*, 43(3):376–395.

Desai, S. and Andrist, L. (2010). Gender scripts and age at marriage in India. *Demography*, 47(3):667–687.

Desai, S., Dubey, A., Joshi, B. L., Sen, M., Shariff, A., and Vanneman, R. (2009). India human development survey: Design and data quality. *IHDS technical paper*, 1:1–27.

Deshpande, A. (2011). *The grammar of caste: Economic discrimination in contemporary India*. Oxford University Press, New Delhi.

Deshpande, A., Goel, D., and Khanna, S. (2018). Bad karma or discrimination? Male–female wage gaps among salaried workers in India. *World Development*, 102:331–344.

Dewan, P., Ray, T., and Roy Chaudhuri, A. (2018). Gender peer effects in high schools: Evidence from India. *Working Paper, Indian Statistical Institute, Delhi*.

Dickson, L. (2010). Race and gender differences in college major choice. *The Annals of the American Academy of Political and Social Science*, 627(1):108–124.

Diekman, A. B., Brown, E. R., Johnston, A. M., and Clark, E. K. (2010). Seeking congruity between goals and roles: A new look at why women opt out of Science, Technology, Engineering, and Mathematics careers. *Psychological science*, 21(8):1051–1057.

Dolado, J. J., Felgueroso, F., and Jimeno, J. F. (2002). Recent trends in occupational segregation by gender: A look across the Atlantic. *Available at SSRN: https://ssrn.com/abstract=320108*.

Dolton, P. J. and Vignoles, A. (2002). The return on post–compulsory school Mathematics study. *Economica*, 69(273):113–142.

Doss, C. (2013). Intrahousehold bargaining and resource allocation in developing countries. *The World Bank Research Observer*, 28(1):52–78.

Dribe, M. and Lundh, C. (2012). Intermarriage, value context and union dissolution: Sweden 1990–2005. *European Journal of Population*, 28(2):139–158.

Duflo, E. (2000). Grandmothers and grandaughters: The effects of old age pension on child health in South Sfrica. *Working Paper No. 00-05, Department of Economics, MIT.*

Duflo, E. and Udry, C. (2004). Intrahousehold resource allocation in Côte d'Ivoire: Social norms, separate accounts and consumption choices. *NBER Working Paper No. 10498*.

Dugar, S., Bhattacharya, H., and Reiley, D. (2012). Can't Buy Me Love? A field experiment exploring the trade-off between income and caste-status in an Indian matrimonial market. *Economic Inquiry*, 50(2):534–550.

Dumont, L. (1980). *Homo hierarchicus: The caste system and its implications*. University of Chicago Press, Chicago and London.

Duraisamy, M. and Duraisamy, P. (2016). Gender wage gap across the wage distribution in different segments of the Indian labour market, 1983–2012: Exploring the glass ceiling or sticky floor phenomenon. *Applied Economics*, 48(43):4098–4111.

Eccles, J. S. and Wang, M.-T. (2016). What motivates females and males to pursue careers in Mathematics and Science? *International Journal of Behavioral Development*, 40(2):100–106.

England, P. and Li, S. (2006). Desegregation stalled: The changing gender composition of college majors, 1971-2002. *Gender & Society*, 20(5):657–677.

Fairlie, R. W. (1999). The absence of the African-American owned business: An analysis of the dynamics of self-employment. *Journal of Labor Economics*, 17(1):80–108.

Fairlie, R. W. (2005). An extension of the Blinder-Oaxaca decomposition technique to logit and probit models. *Journal of Economic and Social Measurement*, 30(4):305–316.

Fairlie, R. W., Hoffmann, F., and Oreopoulos, P. (2014). A community college instructor like me: Race and ethnicity interactions in the classroom. *American Economic Review*, 104(8):2567–91.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, 17(4):347–388.

Fink, G., McConnell, M., and Vollmer, S. (2014). Testing for heterogeneous treatment effects in experimental data: False discovery risks and correction procedures. *Journal of Development Effectiveness*, 6(1):44–57.

Fischer, S. (2017). The downside of good peers: How classroom composition differentially affects men's and women's STEM persistence. *Labour Economics*, 46:211–226.

Fletschner, D., Anderson, C. L., and Cullen, A. (2010). Are women as likely to take risks and compete? Behavioural findings from Central Vietnam. *The Journal of Development Studies*, 46(8):1459–1479.

Fortin, N., Lemieux, T., and Firpo, S. (2011). Decomposition methods in economics. In *Handbook of labor economics*, volume 4, pages 1–102. Elsevier.

Fouad, N. A., Singh, R., Fitzpatrick, M. E., and Liu, J. P. (2011). Stemming the tide: Why women leave engineering. *University of Wisconsin-Milwaukee, Final report from NSF Award*, 827553.

Friedberg, L. and Webb, A. (2006). Determinants and consequences of bargaining power in households. *NBER Working Paper No. w12367*.

Friedman-Sokuler, N. and Justman, M. (2016). Gender streaming and prior achievement in high school science and mathematics. *Economics of Education Review*, 53:230–253.

Frijters, P., Islam, A., and Pakrashi, D. (2019). Heterogeneity in peer effects in random dormitory assignment in a developing country. *Journal of Economic Behavior & Organization*, 163:117–134.

Frijters, P., Islam, A., Pakrashi, D., et al. (2017). Can we select the right peers in Indian education? Evidence from Kolkata. *Discussion paper no. 39/16, Department of Economics, Monash University*.

Fryer, R. G. (2007). Guess who's been coming to dinner? Trends in interracial marriage over the 20th century. *The Journal of Economic Perspectives*, 21(2):71–90.

Fryer Jr, R. G. and Levitt, S. D. (2009). An empirical analysis of the gender gap in Mathematics. *NBER Working Paper No. w15430*.

Fu, X. and Heaton, T. B. (2008). Racial and educational homogamy: 1980 to 2000. *Sociological Perspectives*, 51(4):735–758.

Furtado, D. (2012). Human capital and interethnic marriage decisions. *Economic Inquiry*, 50(1):82–93.

Gadgil, M. and Rao, P. S. (1994). A system of positive incentives to conserve biodiversity. *Economic and Political Weekly*, 29(32):2103–2107.

García-Brazales, J. et al. (2020). Caste in Class: Evidence from peers and teachers. *CEMFI Working Paper No. wp2020_2018*.

Gautam, M. (2015). Gender, subject choice and higher education in india: Exploring 'choices' and 'constraints' of women students. *Contemporary Education Dialogue*, 12(1):31–58.

Gevrek, E. Z., Gevrek, D., and Gupta, S. (2013). Culture, intermarriage, and immigrant women's labor supply. *International Migration*, 51(6):146–167.

Giuliano, P. (2020). Gender and culture. *NBER Working Paper No. w27725*.

Gneezy, U., Leonard, K. L., and List, J. A. (2009). Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77(5):1637–1664.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3):1049–1074.

Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.

Griffith, A. L. and Rask, K. N. (2014). Peer effects in higher education: A look at heterogeneous impacts. *Economics of Education Review*, 39:65–77.

Grogger, J. and Eide, E. (1995). Changes in college skills and the rise in the college wage premium. *Journal of Human Resources*, Apr(1):280–310.

Grossbard, S. A., Gimenez-Nadal, J. I., Molina, J. A., et al. (2014). Racial intermarriage and household production. *Review of Behavioral Economics*, 1(4):295–347.

Guimarães, C. R. F. F. and Silva, J. R. (2016). Pay gap by gender in the tourism industry of Brazil. *Tourism Management*, 52:440–450.

Gullickson, A. (2006). Education and black-white interracial marriage. *Demography*, 43(4):673–689.

Gylfason, T. and Zoega, G. (2003). Education, social equality and economic growth: A view of the landscape. *CESifo Economic Studies*, 49(4):557–579.

Haddad, L. and Hoddinott, J. (1995). Does female income share influence household

expenditures? Evidence from Côte d'Ivoire. *Oxford Bulletin of Economics and Statistics*, 57(1):77–96.

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, pages 351–388.

Hanushek, E. A., Kain, J. F., Markman, J. M., and Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5):527–544.

Hanushek, E. A. and Woessmann, L. (2010). Education and economic growth. In *Economics of Education*, pages 60–67. Academic Press, Elsevier.

Hardies, K., Breesch, D., and Branson, J. (2013). Gender differences in overconfidence and risk taking: Do self-selection and socialization matter? *Economics Letters*, 118(3):442–444.

Hegewisch, A. and Hartmann, H. (2014). Occupational segregation and the gender wage gap: A job half done. *Cornell University, ILR School*.

Hegewisch, A., Liepmann, H., Hayes, J., and Hartmann, H. (2010). Separate and not equal? Gender segregation in the labor market and the gender wage gap. *IWPR Briefing Paper*, 377:1–16.

Hnatkovska, V., Lahiri, A., and Paul, S. (2012). Castes and labor mobility. *American Economic Journal: Applied Economics*, 4(2):274–307.

Hoffmann, F. and Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on college achievement. *Journal of Human Resources*, 44(2):479–494.

Hoxby, C. (2000). Peer effects in the classroom: Learning from gender and race variation. *NBER Working paper no. w7867*.

Hoxby, C. M. and Weingarth, G. (2005). Taking race out of the equation: School reassignment and the structure of peer effects. *Working paper no. 7867*.

Hunt, J. (2016). Why do women leave science and engineering? *ILR Review*, 69(1):199–226.

Hwang, S.-S., Saenz, R., and Aguirre, B. E. (1995). The SES selectivity of interracially married Asians. *International Migration Review*, 29(2):469–491.

Imberman, S. A., Kugler, A. D., and Sacerdote, B. I. (2012). Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. *American Economic Review*, 102(5):2048–82.

Jain, T., Mukhopadhyay, A., Prakash, N., and Rakesh, R. (2018). Labor market effects of high school Science majors in a high STEM economy. *IZA Discussion Paper No. 11934*.

Jakobsson, N. (2012). Gender and confidence: Are women underconfident? *Applied Economics Letters*, 19(11):1057–1059.

James, E., Alsalam, N., Conaty, J. C., and To, D.-L. (1989). College quality and future earnings: Where should you send your child to college? *The American Economic Review*, 79(2):247–252.

Jann, B. (2006). Fairlie: Stata module to generate nonlinear decomposition of binary outcome differentials.

Jann, B. (2008). The Blinder–Oaxaca decomposition for linear regression models. *The Stata Journal*, 8(4):453–479.

Jeffrey, C. (2002). Caste, Class, and Clientelism: A political economy of everyday corruption in rural North India. *Economic Geography*, 78(1):21–41.

Joshee, R. (2008). Citizenship education in India: From colonial subjugation to radical possibilities. In *The Sage Handbook of Education for Citizenship and Democracy*, pages 175–188. Sage Publications, California.

Justman, M. and Méndez, S. J. (2018). Gendered choices of STEM subjects for matriculation are not driven by prior differences in mathematical achievement. *Economics of Education Review*, 64:282–297.

Kahn, S. and Ginther, D. (2017). Women and STEM. *NBER Working Paper No. w23525*.

Kalmijn, M. (1998). Intermarriage and homogamy: Causes, patterns, trends. *Annual Review of Sociology*, 24(1):395–421.

Kalmijn, M. (2010). Consequences of racial intermarriage for children's social integration. *Sociological Perspectives*, 53(2):271–286.

Kalmijn, M. (2015). The children of intermarriage in four European countries: Implications for school achievement, social contacts, and cultural values. *The Annals of the American Academy of Political and Social Science*, 662(1):246–265.

Kalmijn, M., De Graaf, P. M., and Janssen, J. P. (2005). Intermarriage and the risk of divorce in the Netherlands: The effects of differences in religion and in nationality, 1974–94. *Population Studies*, 59(1):71–85.

Kaur, R. (2010). Khap panchayats, sex ratio and female agency. *Economic and Political Weekly*, 45(23):14–16.

Khanna, S. (2012). Gender wage discrimination in India: Glass ceiling or sticky floor? *Working Paper no 214, Centre for Development Economics, Delhi School of Economics*.

Kim, C., Kim, M. K., Lee, C., Spector, J. M., and DeMeester, K. (2013). Teacher beliefs and technology integration. *Teaching and teacher education*, 29:76–85.

Kugler, A. D., Tinsley, C. H., and Ukhaneva, O. (2017). Choice of majors: Are women really different from men? *NBER Working Paper No. w23735*.

Landaud, F., Ly, S. T., and Maurin, É. (2020). Competitive schools and the gender gap in the choice of field of study. *Journal of Human Resources*, 55(1):278–308.

Lavy, V. and Schlosser, A. (2011). Mechanisms and impacts of gender peer effects at school. *American Economic Journal: Applied Economics*, 3(2):1–33.

Lordan, G. and Pischke, J.-S. (2016). Does Rosie like riveting? Male and female occupational choices. *NBER Working Paper No. w22495*.

Luke, N. and Munshi, K. (2006). New roles for marriage in urban Africa: Kinship networks and the labor market in Kenya. *The Review of Economics and Statistics*, 88(2):264–282.

Lundberg, S. J., Pollak, R. A., and Wales, T. J. (1997). Do husbands and wives pool their resources? Evidence from the United Kingdom child benefit. *Journal of Human Resources*, 32(3):463–480.

Mander, H. and Prasad, G. (2014). India exclusion report 2013-14. *Books for Change, Bangalore, India. http://www. indianet. nl/pdf/IndiaExclusionReport2013-2014. pdf.*

Marsh, H. W., Martin, A. J., and Cheng, J. H. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology*, 100(1):78.

Mattoo, M. I. (2013). Career choices of secondary students with special reference to gender, type of stream and parental education. *Research on Humanities and Social Sciences*, 3(20):55–61.

Maurer-Fazio, M., Rawski, T. G., and Zhang, W. (1999). Inequality in the rewards for holding up half the sky: Gender wage gaps in China's urban labour market, 1988-1994. *The China Journal*, (41):55–88.

Mayer, A. K. (2011). Does education increase political participation? *The Journal of Politics*, 73(3):633–645.

McEwan, P. J. (2003). Peer effects on student achievement: Evidence from Chile. *Economics of Education Review*, 22(2):131–141.

McGuinness, S. (2003). University quality and labour market outcomes. *Applied Economics*, 35(18):1943–1955.

Meng, X. and Gregory, R. G. (2005). Intermarriage and the economic assimilation of immigrants. *Journal of Labor Economics*, 23(1):135–174.

Menon, N. and Van der Meulen Rodgers, Y. (2009). International trade and the gender wage gap: New evidence from India's manufacturing sector. *World Development*, 37(5):965–981.

Merton, R. K. (1941). Intermarriage and the social structure: Fact and theory. *Psychiatry*, 4(3):361–374.

Milewski, N. and Kulu, H. (2014). Mixed marriages in Germany: A high risk of divorce for immigrant-native couples. *European Journal of Population*, 30(1):89–113.

Montmarquette, C., Cannings, K., and Mahseredjian, S. (2002). How do young people choose college majors? *Economics of Education Review*, 21(6):543–556.

Moon, V. and Narke, H. (2014a). *Dr. Babasaheb Ambedkar Writings and Speeches Vol. II*. Dr. Ambedkar Foundation, Ministry of Social Justice and Empowerment, Govt. of India.

Moon, V. and Narke, H. (2014b). *Dr. Babasaheb Ambedkar Writings and Speeches Vol. III*. Dr. Ambedkar Foundation, Ministry of Social Justice and Empowerment, Govt. of India.

Mosse, D. (2018). Caste and development: Contemporary perspectives on a structure of discrimination and advantage. *World Development*, 110:422–436.

Munshi, K. (2016). Caste networks in the modern Indian economy. In *Development in India*, pages 13–37. Springer.

Munshi, K. (2017). Caste and the Indian economy. *Cambridge Working Paper Series No. 1759*.

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

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, 15(6):1044–1045.

National Policy on Education, . (1968). *Government of India*.

National Policy on Education, . (1986). *Department of Education, Ministry of Human Resource Development, Government of India*.

Neumark, D. (1988). Employers' discriminatory behavior and the estimation of wage discrimination. *Journal of Human Resources*, pages 279–295.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122(3):1067–1101.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, pages 693–709.

Oaxaca, R. L. and Ransom, M. R. (1994). On discrimination and the decomposition of wage differentials. *Journal of Econometrics*, 61(1):5–21.

O'Neill, J. (2003). The gender gap in wages, circa 2000. *American Economic Review*, 93(2):309–314.

Oosterbeek, H. and Van Ewijk, R. (2014). Gender peer effects in university: Evidence from a randomized experiment. *Economics of Education Review*, 38:51–63.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204.

Paredes, V. (2014). A teacher like me or a student like me? Role model versus teacher bias effect. *Economics of Education Review*, 39:38–49.

Park, G., Lubinski, D., and Benbow, C. P. (2007). Contrasting intellectual patterns predict creativity in the Arts and Sciences: Tracking intellectually precocious youth over 25 years. *Psychological Science*, 18(11):948–952.

Patacchini, E., Rainone, E., and Zenou, Y. (2017). Heterogeneous peer effects in education. *Journal of Economic Behavior & Organization*, 134:190–227.

Patnaik, A., Wiswall, M. J., and Zafar, B. (2020). College majors. *NBER Working Paper No. w27645*.

Penner, A. M. and Paret, M. (2008). Gender differences in Mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, 37(1):239–253.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *BMJ*, 316(7139):1236–1238.

Phipps, S. A. and Burton, P. S. (1998). What's mine is yours? The influence of male and female incomes on patterns of household expenditure. *Economica*, 65(260):599–613.

Pirinsky, C. (2013). Confidence and economic attitudes. *Journal of Economic Behavior & Organization*, 91:139–158.

Plantenga, J., Remery, C., et al. (2006). The gender pay gap. Origins and policy responses. A comparative review of thirty European countries. *European Commission, Directorate-General for Employment, Social Affairs and Equal Opportunities*.

Pope, D. and Sydnor, J. (2010). A new perspective on stereotypical gender differences in test scores. *Journal of Economic Perspectives*, 24(95):108.

Prakasam, G. R., Mukesh, and Gopinathan, R. (2019). Enrolment by academic discipline in higher education: Differential and determinants. *Journal of Asian Business and Economic Studies*, 26(2):265–285.

Przeworski, A. and Limongi, F. (1997). Modernization: Theories and facts. *World Politics*, 49(02):155–183.

Qian, Z. (1997). Breaking the racial barriers: Variations in interracial marriage between 1980 and 1990. *Demography*, 34(2):263–276.

Qian, Z. and Lichter, D. T. (2001). Measuring marital assimilation: Intermarriage among natives and immigrants. *Social Science Research*, 30(2):289–312.

Rao, G. (2019). Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools. *American Economic Review*, 109(3):774–809.

Rapoport, B. and Thibout, C. (2018). Why do boys and girls make different educational choices? The influence of expected earnings and test scores. *Economics of Education Review*, 62:205–229.

Reilly, B., Dutta, P. V., et al. (2005). The gender pay gap and trade liberalisation: Evidence for India. *Poverty Research Unit at Sussex Working Paper*, 32.

Rendall, M. (2013). Structural change in developing countries: Has it decreased gender inequality? *World Development*, 45:1–16.

Riegle-Crumb, C. and King, B. (2010). Questioning a white male advantage in STEM: Examining disparities in college major by gender and race/ethnicity. *Educational Researcher*, 39(9):656–664.

Riegle-Crumb, C., King, B., Grodsky, E., and Muller, C. (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time. *American Educational Research Journal*, 49(6):1048–1073.

Right of Children to Free and Compulsory Education Act, . (2009). *Legislative Department, Ministry of Law and Justice, India*.

Ross, C. E. and Wu, C.-l. (1995). The links between education and health. *American Sociological Review*, pages 719–745.

Rubery, J. (1992). Pay, gender and the social dimension to Europe. *British Journal of Industrial Relations*, 30(4):605–621.

Ruhm, C. J. (2003). Good times make you sick. *Journal of health economics*, 22(4):637–658.

Rukmini, S. (13 November, 2014). Just 5% of Indian marriages are inter caste: survey. *The Hindu*. *https://www.thehindu.com/data/just-5-per-cent-of-.indian-marriages-are-intercaste/article6591502.ece*.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*, volume 3, pages 249–277. Elsevier.

Sacerdote, B. (2014). Experimental and quasi-experimental analysis of peer effects: Two steps forward? *Annual Review of Economics*, 6(1):253–272.

Sadker, M. and Sadker, D. (1986). Sexism in the classroom: From grade school to graduate school. *The Phi Delta Kappan*, 67(7):512–515.

Sahoo, S. and Klasen, S. (2020). Gender segregation in education and its implications for labour market outcomes: Evidence from India. *Demography (forthcoming)*.

Sarsons, H. and Xu, G. (2015). Confidence men? Gender and confidence: Evidence among top economists. *Harvard University, Department of Economics, Littauer Center*, 14:1–26.

Schulz, J. (2019). Kin-Networks and Institutional Development. *Available at SSRN: https://ssrn.com/abstract=2877828*.

Sekhri, S. (2011). Affirmative action and peer effects: Evidence from caste based reservation in general education colleges in India. *University of Virginia. Retrieved February*, 23:2014.

Sekhri, S. (2012). Does academic peer quality promote solidarity? Evidence from caste based peer effects in India. *Mimeo*.

Sen, A., Goutam, P., and Chatterjee, C. (2012). Peer effects in graduate education: Evidence from India. *Working paper*.

Shah, G. (1985). Caste, class and reservation. *Economic and Political Weekly*, 20(3):132–136.

Srinivas, M. N. (1962). *Caste in modern India and other essays*. Bombay and London: Asia Publication House.

Stephan, W. G. and Stephan, C. W. (1991). Intermarriage: Effects on personality, adjustment, and intergroup relations in two samples of students. *Journal of Marriage and the Family*, 53(1):241–250.

Stopnitzky, Y. (2012). The bargaining power of missing women: Evidence from a sanitation campaign in India. *Available at SSRN 2031273*.

Streiner, D. L. (2015). Best (but oft-forgotten) practices: The multiple problems of multiplicity - whether and how to correct for many statistical tests. *The American Journal of Clinical Nutrition*, 102(4):721–728.

Tellhed, U., Bäckström, M., and Björklund, F. (2017). Will I fit in and do well? The importance of social belongingness and self-efficacy for explaining gender differences in interest in STEM and HEED majors. *Sex Roles*, 77(1-2):86–96.

Thomas, D. (1990). Intra-household resource allocation: An inferential approach. *Journal of Human Resources*, 25(4):635–664.

Thomas, D. (1994). Like father, like son; like mother, like daughter: Parental resources and child height. *Journal of Human Resources*, 29(4):950–988.

Thorat, S. and Newman, K. S. (2007). Caste and economic discrimination: Causes, consequences and remedies. *Economic and Political Weekly*, 42(41):4121–4124.

Tucker, S. and Sahgal, G. (2012). 25% Reservation under the RTE: Unpacking the rules in PAISA states forward? *Accountability Initiative, AI Policy Briefs*.

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

Vaid, D. (2014). Caste in contemporary India: Flexibility and persistence. *Annual Review of Sociology*, 40:391–410.

Valla, J. M. and Ceci, S. J. (2014). Breadth-based models of women's underrepre-

sentation in stem fields: An integrative commentary on Schmidt (2011) and Nye et al.(2012). *Perspectives on Psychological Science*, 9(2):219–224.

Velaskar, P. (2012). Education for liberation: Ambedkar's thought and Dalit women's perspectives. *Contemporary Education Dialogue*, 9(2):245–271.

Wai, J., Cacchio, M., Putallaz, M., and Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30 year examination. *Intelligence*, 38(4):412–423.

Wang, M.-T. and Degol, J. L. (2017). Gender gap in Science, Technology, Engineering, and Mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, 29(1):119–140.

Wasserman, M. (2015). Hours constraints, occupational choice and fertility: Evidence from medical residents. *MIT Working Paper*.

Webber, D. A. (2016). Are college costs worth it? How ability, major, and debt affect the returns to schooling. *Economics of Education Review*, 53:296–310.

Winston, G. and Zimmerman, D. (2004). Peer effects in higher education. In *College choices: The economics of where to go, when to go, and how to pay for it*, pages 395–424. University of Chicago Press.

Wiswall, M. and Zafar, B. (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies*, 82(2):791–824.

Wong, H.-P. C. (2014). The effects of endogamous marriage on family outcomes: Evidence from exogenous variation in immigrant flows during 1900-1930 in the United States. *Department of Economics, West Virginia University*.

Wu, A. H. (2017). Gender stereotyping in academia: Evidence from economics job market rumors forum. *Working Paper 2017-09, Center for Health and Wellbeing, Woodrow Wilson School of Public and International Affairs, Princeton University*.

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

Zene, C. (2018). Justice for the excluded and education for democracy in BR Ambedkar and A. Gramsci. *Rethinking Marxism*, 30(4):494–524.