How English exposure in primary education shapes STEM choice: Lessons from India

Bharti Nandwani

Indira Gandhi Institute of Development Research

Mumbai, India.

and

Gitanjali Sen

ORCID ID: 0000-0001-6947-3664

Corresponding Author

Shiv Nadar Institution of Eminence, Gautam Buddh Nagar, UP 201314, India.

Email: gitan_dutta@hotmail.com and Gitanjali.sen@snu.edu.in

Acknowledgement: The authors acknowledge helpful comments from the seminar participants of the Asia Meeting of the Econometric Society held at IIT Delhi, 2024, BITS-Pilani Hyderabad, 2024, and AFDEV conference at Aix Marseille University, France in 2024.

Importance of English as a Second Language in Higher Education: Learning from a Natural Experiment in India

Abstract

English, the world's third most spoken language, is of global importance, especially in former British colonies, which are destinations for outsourced jobs. In India, with its many languages, English serves as the lingua franca, providing labor market advantages. Despite India's emphasis on vernacular languages in school education, English remains the primary medium of instruction in higher education, particularly for STEM subjects. Therefore, learning English in schools is expected to benefit students pursuing STEM. Through a language policy variation in West Bengal, India, we investigate the impact of English language exposure in primary school on subject choice in higher education. Our analysis shows that the exposed students have higher chances of enrolment in technical or STEM subjects in higher education by 5 to 7.6 percentage points. The chances of dropping out from higher education are reduced by 7.2 percentage points on average. The positive effects of English exposure on STEM selection are less pronounced among females. Improved English speaking skills seem to be one of the mechanisms, and we do not find any effect on the overall quality of education.

Keywords: Education, STEM, English, Subject Choice, Language Policy, India.

JEL Codes: I25; I28; O53

1. Introduction

The growing importance of English as a global language has gained attention, especially because it is the third most spoken language worldwide. English skills have become increasingly valuable in countries that were formerly British colonies and where many jobs have been outsourced in recent decades. In places like India, mastering English has been shown to lead to better job opportunities and higher earnings (Azam et al., 2013; Chakraborty and Bakshi, 2016; Munshi and Rosenzweig, 2005; Shastry, 2012). Research by Jain et al. (2019) establishes that proficient spoken English can result in 32% higher earnings. Notably, Jain (2019) identifies a lack of English-speaking skills as a significant barrier to employment and consistent income in India.

The English colonial system heavily influenced the modern education system of India because India was a British colony for two centuries until its independence in 1947. In 1885, Lord Dalhousie accepted Wood's recommendation to adopt vernacular languages for instruction at the primary and secondary school levels and the English language for instruction at the higher education level. Henceforth, the undergraduate classes in most Indian universities have been taught in English for more than a century[1]. In a multilingual country like India, which has 22 scheduled languages and more than 122 regional languages, English is an important *lingua franca.* English is particularly important in districts where a language mismatch between the language spoken by the majority and the medium of instruction (that is, the official State language) is found to generate poor education outcomes (Jain, 2017). Although the state governments of India have started to use the "local" language as the medium of instruction in public schools, even now, the medium of instruction in the majority of the higher education institutions (henceforth, HEI)[2] is English and more so for the Science, Technology, Engineering and Math (STEM) subjects.

Therefore, studying English in school is expected to benefit STEM students in higher education compared to those in other subjects, as most non-STEM subjects are taught in local

[1] Except for the Osmania University in Hyderabad, where it was taught in Urdu.

[2] All undergraduate degree colleges, postgraduate degree colleges, universities, or autonomous institutes that can offer degrees or diplomas and require a minimum of twelve years of high school education are defined as "higher education institutions" or HEI throughout the paper.

languages in public HEIs. We assess how early exposure to formal English language as a subject impacts students' subject choices in higher education. In our study conducted in West Bengal, an eastern state of India, we utilize a language policy change that introduced English as a compulsory subject in primary classes in public schools to determine the impact of early exposure to English language education. We use the birth year and state of residence of school children to estimate the impact of treatment in a classic Difference in Difference (DID) setup where five neighboring states are used as the control states.

Our analysis using the Ordinary Least Squares (OLS) model suggests that students who are exposed to English as a second language subject from primary school have a greater chance of enrolling in technical and STEM subjects, on average. The chances of choosing technical and STEM-related fields are higher by 7.6 and 5.1 percentage points, respectively, as compared to those who are not exposed. We also observe a 7.2 percentage point decline in college dropout due to English exposure. Interestingly, the positive effects of English exposure on selecting STEM subjects are more pronounced among male students. We also find that the effects are pronounced when children are fully exposed in the primary classes starting from 1st grade. We find suggestive evidence that the command over the English language is one channel through which students may have envisioned the advantage of studying STEM in later years.

Our research contributes to three primary strands of literature. Firstly, it adds to the understanding of how exposure to English as a second language influences job opportunities globally (Bleakley and Chin, 2010; Chakraborty and Bakshi, 2016; Chatterjee, 2018; Kossoudji, 1988). Secondly, it sheds light on the significance of English in countries with colonial history like India (Azam et al., 2013; Chakraborty and Bakshi, 2016; Munshi and Rosenzweig, 2005). Thirdly, it explores the relationship between investment in English language skills and educational outcomes (Jain, 2017; Kingdon, 2020; Maple & Stage, 1991; Nandwani and Sen, 2023; Roy, 2004).

Our study builds upon the work of Chakraborty and Bakshi (2016), who analyzed the impact of removing English education from primary schools in West Bengal. They found that English proficiency is correlated with higher earnings and employment in high-ranked jobs. Other studies on the same policy, such as those by Roy (2004) and Nandwani and Sen (2024), observed

increased private investment in education among students not exposed to English at the primary level. Nandwani and Sen (2024) revealed that removing English led to more enrolment in private schools and increased spending on private tuition for students in public schools. These studies establish the potential returns arising from English language skills. However, the literature in this area misses the important link towards higher labor market returns, which works through subject choice in college. We are not aware of studies that examine the effects of English language policy on subject choices in tertiary education. Our research extends the findings of Nandwani and Sen (2024) by investigating the long-term impacts of language policy on subject selection in higher education, highlighting its enduring effects.

In the next section, we delve into the context of our study in detail. This is followed by an explanation of the data we utilize, how we construct variables, and the method we use to identify patterns in section 3. Section 4 presents the key findings of our research and discusses the heterogeneity of estimates, test of robustness, and potential mechanism. Finally, in section 5, we conclude the findings.

2. **Context**

The English language policy in West Bengal has a long history. English was taught as a second language in all vernacular medium government schools in West Bengal from the first grade. However, when the Left Front government came into power in 1977, they decided to remove English as a second language subject from primary classes (grades 1 to 5) of public schools, and was to be taught only in grade 6 of secondary school in 1983. English being a foreign language, this change in curriculum was expected to reduce the dropout rate in West Bengal (Bagchi, 1981). However, students began learning the English alphabet only in secondary school, leading to a complete revision of the syllabus at the secondary and higher secondary levels. As a result, the English language proficiency of this cohort during their high school years was lower than that of those who started learning English in the first grade.

However, concerns were raised by various segments of society, and there was evidence of these cohorts facing challenges in the job market. Criticism of the English removal policy

escalated in the late 1990s, particularly from opposition parties, and there was a growing demand for English education from the middle class. Consequently, the government gradually began to reverse the policy. English was reintroduced in the 3rd grade in 1999 and from the 1st grade in 2004. Using the English re-introduction policy change as a natural experiment, we create a binary variable to measure English exposure. This variable takes a value of one for the cohort that studied English at the primary level (hence the exposed cohort) and takes a value of zero for the cohort that did not study English at the primary level. Since this policy pertained to West Bengal, we use neighboring states, namely Assam, Bihar, Orissa, Jharkhand, and Sikkim, as control states in our difference-in-difference analysis.

Figure A1 in the appendix tabulates the reasons behind school choice by the currently studying private school students, plotted separately for West Bengal and other states[3] in the year 2014[4]. It demonstrates that 17 percent of high school students in West Bengal and 6 percent in the rest of India opted for private schools primarily for English medium instruction in 2014, suggesting high demand for learning English language in West Bengal. Figure A2 shows a rising trend in enrollment in English medium HEIs (as well as schools) as a proportion of total enrollment over time. Additionally, the percentage of enrollment in English medium is much higher in higher education as compared to schools both in West Bengal and the rest of the country.

A preliminary examination of the latest education round of the NSS data from 2017-18 suggests that in West Bengal, the medium of instruction in higher education, including graduate and post-graduate courses, is divided between English and Bengali (vernacular) languages. While humanities subjects such as Political Science, Philosophy, Education, History, and Sociology are taught in the vernacular language in government colleges, science subjects are primarily taught in English.

Figure A3 shows that STEM enrollment as a percentage of total enrollment has consistently been higher in English medium HEIs in both West Bengal and the rest of India. Additionally, the rate of STEM enrollment has been increasing faster in English medium HEIs compared to vernacular medium HEIs. In the most recent data from 2017-18, STEM enrollment

[3] We include all other states in India in this figure.
[4] This is the earliest round of data for which the reasons for choosing private schools were asked.

in English medium colleges has surpassed 60 percent in West Bengal, while it is around 52 percent in the rest of India.

The English introduction policy is likely to improve the chances of choosing STEM subjects for the exposed cohort in college as an early introduction to English improves command over the language. On the other hand, cohorts that are not exposed to the policy may worry about struggling academically in higher education without a strong English background. Additionally, students with improved English skills from their school years may be less likely to drop out of college, particularly STEM subjects. Thus, the hypothesis we test is that the exposed cohort has a higher likelihood of studying subjects like medicine, engineering, management, and basic science, which are mainly taught in English in Indian colleges and universities. Additionally, we test if being exposed to English early in school leads to lower dropouts from higher education where a significant number of subjects are taught in English.

3. **Data and Methods**

3.1 Data, sample and outcomes

We analyze data from several rounds of the National Sample Survey (NSS), a large household survey conducted by the Ministry of Statistics and Program Implementation (MOSPI). These surveys gather detailed socio-demographic information from around 0.6 million individuals in each round and are carried out approximately every five years. Additionally, MOSPI conducts education-specific NSS rounds within this five-year period (known as education rounds) to collect detailed information about education from surveyed household members. Our study focuses on the education rounds conducted in 1995-96 (52nd round), 2007-08 (64th round), 2014 (71st round), and 2017-18 (75th round). It is important to note that NSS rounds are designed as cross-sections and do not survey the same individuals across rounds. We use a cross-section of these four rounds of NSS data. Specifically, we focus on West Bengal and its neighboring states: Assam, Bihar, Jharkhand, Orissa, and Sikkim. Within these states, our sample comprises individuals currently enrolled in higher education (undergraduate and beyond). Using this analytical sample, we construct the following outcome variables:

Technical education is a dummy variable taking a value of 1 if the individual *i* is currently attending a professional course (like engineering, medicine, agriculture, etc), and 0 otherwise.

STEM is a dummy variable taking a value of 1 if the individual *i* is currently attending engineering, medicine or Information Technology (IT) related courses, and 0 otherwise. This is therefore a subset of the first variable.

Dropout – For this variable, we restrict our sample to individuals between the ages of 17 and 25 who may or may not be currently pursuing higher education. This variable takes a value 1 if the individual *i* dropped out from a higher education course after completing higher secondary school, and 0 otherwise.

Table A1 in appendix presents summary statistics and shows that the likelihood of pursuing technical degrees, particularly engineering and medicine is higher in West Bengal as compared to other neighboring states. The dropout rates are lower in the state of West Bengal.

The five states that share a border with the state of West Bengal (see appendix Figure A4) qualify as a control group of states because of the linguistic history, the history of state formation in post-independent India, and the consequent education policies. We elaborate on these in the next few paragraphs.

First, the East India Company expanded its territorial acquisitions, beginning with the Bengal province in 1757 until 1857, when the territory came under the direct control of the British Empire through a mutiny. Hence, this region, the Bengal Presidency, was formed long before the reorganization of the state boundaries in independent India in 1956. The six states in our working sample were part of the erstwhile Bengal Presidency. The southern part of India was acquired by the British Empire from the regional rulers between the late seventeenth century and eighteenth century, to be governed as Madras Presidency. The western region, including some islands from Portugal and other regions of Maratha territories, continued to be ruled as the Bombay Presidency from the late nineteenth century. The rest of modern India was acquired from native Kings or Princes at different times. For the ease of administration of these conquered territories and to introduce locales to Western civilization, the British rulers established schools and colleges. They wanted the English language as a part of the curriculum in formal education

(Evans, 2002). Considering this diverse history of exposure to the British Empire and the consequent evolution of the English language as a part of the formal education system, we believe the five states from the erstwhile Bengal presidency form a suitable counterfactual for the treated state of West Bengal.

Second, the linguistic differences were considered when the state administrative boundaries were drawn in 1956 in independent India. The measures of linguistic distance between each pair of Scheduled Indian languages (Table 6 of Jain, 2017) shows that the neighboring states of Orissa, Jharkhand, Bihar, and Assam had the lowest linguistic distance with the Bengali language, which is the state vernacular language of our treated state West Bengal. Following Jain (2017), the higher linguistic distance between a pair means the greater difficulty picking up the other's language. While the distance range for all scheduled languages in India varies between 0 to 13 according to this measure, this distance ranges between 0 to 4 for our sample states (the distance with own language being zero). As we are interested in estimating the effect of exposure to the English language, the difference between difficulty levels of state vernacular languages should be the minimum so that the linguistics distance with English does not confound our estimates.

3.2 Empirical Framework

We present estimates from the following two specifications. First, we estimate the following OLS model with a difference-in-differences specification where the outcome is one of the above-defined variables, and the treatment includes the fully exposed and partially exposed cohorts together:

$$Y_{ithd} = \beta_1 + \beta_2 WB_i * Exposed_t + \gamma X_{ih} + \Gamma_t^1 + \Gamma_d^2 + \epsilon_{ithd}. \qquad (1)$$

Here, Y_{ithd} indicates the outcome for individual i born in year t, belonging to household h, residing in district d. The *Exposed* variable takes a value of 1 for the cohort who could study English at the primary level after the policy reversal in 1999. The cohorts before them, who could not study any English at the primary level (henceforth studied English of a lower quality

when it was introduced at the secondary level), are our control cohort. For this latter group, the variable *Exposed* takes a value zero. Thus, the cohorts born after 1992 are exposed to the policy, and cohorts born between 1973 and 1992 are the control cohort[5].

WB is a dummy variable that takes a value 1 for the individual belonging to the state of West Bengal (treated state) and it takes a value of zero for individuals from the neighboring (control) states. X_{ih} is a vector of individual and household level observable covariates such as, age and sex of the child, household size, the monthly consumption expenditure of the household as an indicator of economic conditions, the indicator capturing whether the household head had at least secondary education, the indicator for the sex of the household head, distance to the nearest school, and indicator capturing whether the individual resides in rural or urban area. The summary statistics of these covariates are reported in Appendix Table A2 for the treatment and control states for the pre-policy cohorts. Γ_t^1 is a vector of birth year fixed effects controlling for time-specific unobservable covariates. Γ_d^2 is a vector of district-fixed effects controlling for district-specific time-invariant unobservable factors that could influence the outcomes. We bootstrap all our standard errors because clustering those at such few cluster levels would lead to a wrong inference (Bertrand et al 2004).

The coefficient of interest is β_2 which captures the effect of studying English from the primary level on subject choice in college or in dropout outcome.

However, the cohorts born between 1993 and 1997 are partially exposed to the policy. So in the next specification, we present estimates separately for the fully exposed and partially exposed cohorts using the same sample:

$$Y_{ithd} = \beta_1 + \beta_2\ WB * Fully\ Exposed_t + \beta_3\ WB_i * Partially\ Exposed_t + \gamma X_{ih} + \Gamma_t^1 + \Gamma_d^2 + \epsilon_{ithd}. \quad (2)$$

The identification strategy rests on the assumption that in the absence of the policy, the outcomes in the exposed and the control groups must follow parallel trends. While it is not

[5] Our estimates are robust to removing cohorts at the margin. We present that result for all outcomes of our preferred specification in appendix Table A4.

possible to empirically test this assumption, we attempt to show in the next section that the outcomes do not show differential trends during the pre-policy period.

The state-level education policies, or the language policies of the neighboring states have not gone through any major change during the treatment period of our interest. The national level education policies such as the Sarva Shiksha Abhiyan (implemented in 2002) leading to infrastructure expansion at the primary level and upper primary level (between the age of 6 to 14), mid-day meal schemes (implemented from 1995), or the Right to Education Act (implemented for increasing access to free public schools for children between age 6 to 14 from the year 2010) are implemented across all states of India. Therefore, these education policies are not expected to systematically affect the control states differently from West Bengal.. However, the level of implementation of these policies could have varied across states or districts which could confound our estimates. Therefore, we also present and compare estimates with and without controlling for district-fixed effects, and our preferred specification controls for district-fixed effects and birth-year fixed effects. Our estimates are not sensitive to the inclusion of district-fixed effects or birth year-fixed effects, which indicates the robustness of our estimates to unobserved location-specific factors or time-variant unobserved factors.

3.3 Pre-existing trends

Before continuing to our main estimation results, we check pre-existing trends in the outcomes between the exposed state and the control group of states. The graphical diagnostics for the three outcomes are presented in Figures 1A to 1C. The vertical line in each figure indicates the birth year after which the cohort got exposed to the English language in primary schools.[6] Before the policy implementation, treated and control states followed a parallel path. We do not observe any significant difference in the mean of outcomes (the left-hand figures) between the treated and control states before the policy implementation.

We also perform a formal test to see if the trajectories are parallel. We extend our baseline model to include variables representing time trends before and after the treatment for

[6] We use the didregress package of Stata 18, where birthyear captures the timeline and states form the group..

both groups of states. The linear-trends model estimates a coefficient for the differences in linear trends prior to treatment, and if that coefficient is 0, the linear pretreatment trends are parallel. The F-stat[7] reported for this test being 0.15 and 0.19, respectively, for technical and STEM outcomes, and 1.52 for the dropout outcome, we do not have evidence to reject the null hypothesis of parallel trends in this case. Graphs based on this test of violation of parallel trends (the right-hand figures) are reported in Figures 1A to 1C.

Figure 1A here

Figure 1B here

Figure 1C here

4. **Results and Discussion**

4.1 Effects of the program on education outcomes

Table 1 here

Table 1 reports the outcomes that are observed after enrolling in higher education. The impact on the subject choice in Technical or STEM subjects are reported in panels A and B, respectively, by estimating specifications (1) and (2). Table 2 reports the outcome regarding the dropout decision, that is, whether the respondent dropped out from education after completing higher secondary education.

For both tables, column 1 reports the OLS estimates from an unconditional model (with level effects only). Column 2 controls for the individual and household level observable covariates as explained earlier, column 3 adds district-fixed effects, and column 4 adds birth year fixed effects, which is our preferred specification.

Column 4 of Panel A shows that the exposed cohort in West Bengal has an 8 percentage point higher likelihood of pursuing technical courses in college than the unexposed cohort in neighboring states when we estimate the average treatment effects of partially and fully exposed

[7] We use ptrends as post-estimation command after didregress using Stata.

cohorts together. However, the estimates of specification 2 show that a fully exposed cohort has a 12 percentage point higher chance of being enrolled in technical subjects. Compared to the baseline mean (0.28) in the sample, this represents a full exposure leads to a 43% increase in the chances of pursuing technical education.

Similarly, column 4 of Panel B suggests a five percentage point increase in the chances of pursuing STEM subjects in higher education on average when we estimate the impact of partially exposed and fully exposed cohorts together. However, a fully exposed cohort has an 8.5 percentage point higher chance of enrolling in STEM subjects (specification 2 of Panel B), which is about 40% increase over the baseline mean of 0.21. In all the final specifications, the p-values are small enough to reject the null hypothesis of equality of coefficients between partial exposure and full exposure.

Table 2 here

Table 2, reports that the exposed cohort has about seven percentage points lower chances of dropout from higher education on average when the impact is estimated for all the cohorts together. Here, we restrict our sample to adults who have completed high school education. For fully exposed cohorts, the chances of dropout are reduced by 4.5 percentage points. When the baseline mean of dropout is as high as 43%, a ten percent lower chance of dropout makes a significant impact on participation in higher education. The impact is higher among partially exposed cohorts for the dropout outcome, which may capture the immediate effect of the exposure to English because this is the cohort for whom the changes started in primary grade.

4.2 Heterogeneous effects of the program

4.2.1 Across Sex

Table 3 here

We restrict our analysis to the average outcomes of partially and fully exposed cohorts in this section for brevity and clarity. Panel A of Table 3 examines if the gain from re-introducing

English subjects in the primary sections is similar across genders. Several studies find that females are under-represented in STEM professions (Bottia et al 2015; Griffith 2010; Kahn and Ginther 2017). So, there is reason to expect that the impact of the policy on the likelihood of females choosing science-related fields in higher education might be less as compared to boys. Our estimates support this hypothesis. Columns 1 and 2 report that exposed females have seven and 7.7 percentage points lower chances of choosing technical and STEM subjects, respectively, than exposed males. However, it is important to note that the impact of English exposure on the exposed male's choice of technical and STEM subjects is 10.5 and 8.2 percentage points higher.

Column 3 reports 7.6 percentage points lower chances of dropout by exposed male, whereas the exposed female do not seem to have significant effect. Overall, we get the evidence of heterogenous effect of the policy, where the English exposure seems to be more beneficial to males.

4.2.2 Across Castes

We also examine if the impact of the policy varied by caste groups - an important marker of social identity in India. The caste groups in the NSS data are broadly categorized into Scheduled Caste (SC), Scheduled Tribes (ST), Other backward Caste (OBC) and general caste category. The first two categories are the most marginalized caste groups in India, followed by the OBC. The general caste category consists of all caste categories that do not fall in the former three categories. Several papers in the Indian context have shown that most welfare programs have benefitted individuals from SC, ST, OBC households less than the general category and therefore this examination is informative. The first and second colunns of Panel B show that the English exposure helped the general category students more in opting for technical or STEM subjects (with 7.9 and 6.5 percentage point higher chances respectively). However, the effect is not so pronounced for other disadvatnged castes on avearge. While we get evidence of heterogenous effects across caste too, the impact on dropout outcomes (in 3rd column of Panel B) seem promising with all caste groups having lower chances of dropout due to English exposure.

4.3 Robustness

4.3.1: Robustness to treatment effect heterogeneity

A recent difference-in-difference literature suggests that if treatment effects are heterogeneous across time or groups, the DID coefficient is unlikely to indicate the average treatment effect (Chaisemartin and Haultfoeuille, 2020; 2022). As our analysis includes multiple years and states, the treatment effect can plausibly be heterogeneous. To address this concern, we estimate the two-stage regression approach proposed by Gardner (2022), which is robust to treatment effect heterogeneity. In the first stage, the outcome is regressed on group and time-fixed effects for untreated observations. In the second stage, the outcome is purged of the group and year-fixed effects and then regressed on the treatment. The two-stage regression results, reported in Appendix Table A3, show that the interaction coefficients are very close to the coefficients estimated in Tables 1 and 2.

4.3.1: Robustness to Randomisation of Treatment

We conduct the test of exact randomization - a standard placebo test to ensure that our estimates are not picking up the spurious correlations. We do so by randomly assigning the treatment status to state-birthyear cohorts and re-estimating regression equation 1 multiple times. Thus, some of the state-cohort pairs that were originally a part of the control group are assigned to the treatment group, and some of the eligible groups have been excluded from the treatment group. We estimate the placebo regressions 1000 times, recording the coefficient and standard errors each time. Ideally, these coefficients should be statistically insignificant, and their magnitude should be smaller relative to our coefficients in Tables 1 and 2. We report the distribution of coefficients in Figure 2. The placebo tests re-reassuringly suggest that most of these coefficients are centered around zero and are very small in magnitudes. Further, we check that our placebo coefficients are significant only 4 to 6% of the time at a 95% significance level for our outcome variables.

Figure 2 here

4.4 Mechanism

Angrist et al. (2008) find that English as a medium of instruction in Puerto Rican School helped in improved English speaking ability among Puerto Rican children. Following this finding, we wanted to explore if the higher ability to converse in English with the exposed children was the mechanism behind the STEM choice. A higher speaking ability may boost their confidence in choosing a subject taught in English. Ideally, we would like to check the writing ability of English, too, as a mechanism. However, the survey question that is asked to all the birth cohorts includes information about whether the individual can converse in English. The categorical variable asks to choose converse with fluency, converse, and cannot converse as three categories. We pool the the first three categories taking a value of 1 and not able to converse as zero.

We use the Indian Human Development Survey data (IHDS 2) from the Round 2 collected in the year 2011-12. The IHDS is household-level nationally representative panel data collected in 2004-05 (Round 1) and 2011-12 (Round 2). This data collects detailed information on the household members about their income, expenditures, standard of living, education, and health. However, for the outcome of our interest, we could not use the panel nature of the data because only IHDS 2 has information on both the treated and control cohorts.

Using the above variable as an outcome, we estimate a specification similar to that of (1), with a few changes in covariate. The covariates include the state dummy, birth year dummy, urban dummy, gender, caste and education of household head, household consumption expenditures per capita, and an indicator of the mother tongue is the same as the state vernacular language.

Table 4 here

Column 1 of Table 4 shows that the exposed cohort in West Bengal has about 3.7 percentage points higher chances of being able to converse in English.

4.5 Potential Concerns

One major concern would be the omitted variables that could affect the outcomes of the treated state differently from those of the selected group of control states over a period of time. As we restrict our analysis to the neighboring states, which are expected to have similarities in demographics, we expect to minimize such bias.

None of our preferred specifications control for time-varying state-specific linear trends arising due to different development outcomes potentially affecting education choices. This may confound our estimates of treatment effect. However, our treated state being a single state, any such covariate would absorb treatment effects too. Therefore, we resort to checking the robustness of our estimates by including covariates in a sequential manner. Our estimates remain robust to the inclusion of district-specific or birthyear-specific changes, which indicates that the unobserved factors varying across time or district or state may not drive the results. In a separate analysis (not presented here), we include district-time trends instead of district-fixed effects and birth year-fixed effects, and all our estimates are robust to that specification[8].

Additionally, we randomly drop one control state at a time and estimate the treatment effects with the remaining sample. Appendix Table A4 presents estimates from all such variations and establishes that our estimates are robust to changes in the composition of the control group. If the state-specific time variant omitted variable would bias our estimates, we would expect our estimates to be affected. We also test the potential of measurement errors in reporting birth years at the margin by excluding a few birth cohorts of the older ones. A few variants of such estimates provide the confidence that our treatment effects are not confounded by other time-varying factors or measurement errors in the reporting of birth year. We present one such result in Appendix Table A5.

What we are able to capture is the length of exposure to English as a second language, but we are unable to differentiate between the length of exposure and the quality of overall teaching. The quality of teaching may vary depending on the availability of qualified instructors, study materials, and resources. Our estimates would be biased if these would vary differently across places and over time. If the quality of teaching had been affected, we would expect the general writing skills to be affected as well. As IHDS data surveys the writing skills of students through

[8] We could not use synthetic control methodology because of a small number of pre-period state-birthyear level observations even while using all Indian states.

some writing tests, and the way the skills are tested has nothing to do with English language skills, we use that variable as a proxy for quality of education. A child is allowed to test their writing skill based on his or her own chosen language. Our estimates in column 2 of Table 4 show no effect on the writing skill of the student in general when she is given a choice to take the test in her own preferred language. This finding establishes that exposure to English may have affected the skills to converse in English but not the overall quality of teaching.

5. Conclusion

English has emerged as a major global language, and numerous studies have shown that proficiency in such a globally significant language can lead to higher earnings in the job market over time. This is especially noticeable in countries with colonial histories, where the former colonizers had already invested in enhancing skills in the colonial language. While the importance of learning one's native language cannot be overlooked, as it is often the best medium of instruction during early childhood, research has also demonstrated that children are capable of learning multiple languages at a young age. In this context, it becomes crucial for policymakers to explore whether it is beneficial for children to start learning English as a foreign language early in life, even if their primary medium of instruction is vernacular.

Using a language policy change in West Bengal, India as a source of exogenous variation, we study the impact of exposure to English as a second language from early school years on the subject choice outcomes at the higher education levels. Since English is the primary language of instruction in STEM subjects at most higher education institutions, we find that children exposed to English as a second language from a young age are more likely to pursue technical and STEM subjects and less likely to drop out from higher education. A particularly interesting aspect of our findings is the varied impact across gender and castes, with females exposed to English showing a lower increase in the likelihood of pursuing technical education. We also show that the policy effects are more pronounced among advantaged caste groups. We find some suggestive evidence that improved English language skills being helpful in their future subject choice.

Our findings hold significant relevance for policymakers, particularly considering India's acknowledgment of the importance of multilingualism in the New Education Policy of 2020. While discussions surrounding language policies in India and elsewhere often center on the medium of instruction, our research provides evidence of the importance of English as a foreign language in multilingual countries. Building upon the work of Nandwani and Sen (2024), which highlights the increased demand for English in India and its impact on private investment in education, our paper extends this research by demonstrating how exposure to the English language from an early age can influence subject choice in higher education. This underscores the broader implications of language policies beyond just the medium of instruction.

References

Acharya, P. 1982. Abolition of English at Primary Level in West Bengal, *Economic and Political Weekly*, 17(4): 124-128, January 23.

Azam, M., Chin, A., & Prakash, N. (2013). The returns to English-language skills in India. Economic Development and Cultural Change, 2(61), 335–367.

Bagchi, J. 1981. Storm over English in West Bengal: A Comment. *Economic and Political Weekly*, 16(22): 993–994.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?." *The Quarterly journal of economics* 119.1 (2004): 249-275.

Bleakley, H., & Chin, A. (2010). Age at arrival, English proficiency, and social assimilation among US immigrants, American Economic Journal. Applied Economics, 2(1), 165–192.

Bottia, Martha Cecilia, et al. "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 (2015): 14-27.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. "Two-way fixed effects estimators with heterogeneous treatment effects." *American economic review* 110.9 (2020): 2964-2996.

De Chaisemartin, Clément, and Xavier d'Haultfoeuille. *Two-way fixed effects and differences-in-differences estimators with several treatments*. No. w30564. National Bureau of Economic Research, 2022.

Chakraborty, T. and Bakshi, S. K. 2016. English language premium: Evidence from a policy experiment in India, *Economics of Education Review*, 50:1–16.

Chatterjee, S. 2018. Do Private Tutors Enhance English Language Ability? Regression Discontinuity Evidence from a Policy Experiment in India. *Bulletin of Economic Research*, 70:139-149. https://doi.org/10.1111/boer.12142

Datta, B. 1981. Storm over English in West Bengal. *Economic and Political Weekly*, 16(14): 619–622. http://www.jstor.org/stable/4369675

Evans, Stephen. 2002. "Macaulay's Minute Revisited: Colonial Language Policy in Nineteenth-Century India." *Journal of Multilingual and Multicultural Development* 23, 4:260–81.

Gardner, John. 2022. "Two-stage differences in differences." *arXiv preprint arXiv:2207.05943* , accessed on 21st April, 2024.

Guha Thakurta, S.N. 1998. Rigmarole over English in West Bengal, *Economic and Political Weekly*, 33(22):1300-1303.

Griffith, Amanda L. "Persistence of women and minorities in STEM field majors: Is it the school that matters?." *Economics of Education Review* 29.6 (2010): 911-922.

Jaekel, N., M. Schurig, M. Florian, and M. Ritter. 2017. From Early Starters to Late Finishers? A Longitudinal Study of Early Foreign Language Learning in School. *Language Learning,* 67(3): 631–664.

Jain, T. (2017). Common Tongue: The Impact of Language on Educational Outcomes. *The Journal of Economic History*, 77(2)4: 73-510. doi:10.1017/S0022050717000481

Jain, Tarun & Maitra, Pushkar & Mani, Subha, 2019. "Barriers to skill acquisition: Evidence *from* English training in India," World Development, Elsevier, vol. 114(C), pages 314-325.

Kahn, Shulamit, and Donna Ginther. *Women and STEM*. No. w23525. National Bureau of Economic Research, 2017.

Kingdon, G. G. 2020. The Private Schooling Phenomenon in India: A Review, *The Journal of Development Studies*, 56(10):1795-1817, DOI: 10.1080/00220388.2020.1715943

Kossoudji, S. A. 1988. English Language Ability and the Labor Market Opportunities of Hispanic and East Asian Immigrant Men, *Journal of Labor Economics*, 6(2):205-228.

Maple, S., & Stage, F. (1991). Influences on the choice of math/science major by gender and ethnicity. American Educational Research Journal, 28(1), 37–60.

Munshi, K., and Rosenzweig, M. 2005. Traditional Institutions Meet the Modern World: Caste, Gender and Schooling Choice in a Globalizing Economy. *American Economic Review*, 96 (4): 1225-1252.

Nandwani, Bhrati and Gitanjali Sen. 2024. Importance of English as a Second Language: Learnings from a Language Experiment in India, working paper.

Nikolov, M. and J. M. Djigunovic. 2006. Recent Research on Age, Second Language Acquisition, and Early Foreign Language Learning. *Annual Review of Applied Linguistics,* 26: 234-260.

Roy, J. (2004). 'The impact of lowering of academic standards on educational outcomes: Evidence from an unusual policy in India', Unpublished Manuscript.

Sen, S. 2015. Abolishing English in schools: implications for higher education in West Bengal. Inter-Asia Cultural Studies 16.2: 269-281.

Shastry, G. K. 2012. Human Capital Response to Globalization: Education and Information Technology in India. The Journal of Human Resources 47(2): 287–330. http://www.jstor.org/stable/23214392.

Table 1: Exposure to English as a second language on HE outcomes if currently Attending

OLS Estimates	(1)	(2)	(3)	(4)
Panel A: Dependent Variable: Technical Education =1, Else 0 (baseline mean: 0.28)				
Specification 1: Combined treatment for partial and full exposure				
WB X Exposed	0.084***	0.049**	0.054***	0.076***
	(0.025)	(0.021)	(0.019)	(0.022)
Specification 2: Separate Dummies for partial and full exposure				
WB X Partially Exposed	0.054**	0.026	0.032	0.057***
	(0.026)	(0.024)	(0.023)	(0.020)
WB X Fully Exposed	0.121***	0.091***	0.095***	0.121***
	(0.031)	(0.026)	(0.029)	(0.026)
p-value: Test of Diff in Coeff	0.017	0.025	0.02	0.006
Observations	8,550	8,549	8,545	8,545
Panel B: Dependent Variable: STEM Education =1, Else 0 (baseline mean: 0.21)				
Specification 1: Combined treatment for partial and full exposure				
WB X Exposed	0.045**	0.016	0.024	0.051***
	(0.019)	(0.019)	(0.021)	(0.018)
Specification 2: Separate Dummies for partial and full exposure				
WB X Partially Exposed	0.017	-0.004	0.004	0.037
	(0.018)	(0.025)	(0.021)	(0.026)
WB X Fully Exposed	0.091***	0.051*	0.061***	0.085***
	(0.026)	(0.029)	(0.023)	(0.029)
p-value: Test of Diff in Coeff	0.013	0.034	0.015	0.009
Observations	8,550	8,549	8,545	8,545
Controls	No	Yes	Yes	Yes
District FE	No	No	Yes	Yes
Birthyear FE	No	No	No	Yes
Exposed Cohort	Yes	Yes	Yes	No
West Bengal	Yes	Yes	No	No

Source: NSS Data. Bootstrapped standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1. Panel 1 has technical education and panel 2 has STEM education as outcomes. We report the p-values from the test of the null hypothesis of equality between coefficients of partial exposure and full exposure.

Table 2: Exposure to English as a second language on HE outcomes of Dropout

OLS Estimates	(1)	(2)	(3)	(4)
Dependent Variable: Dropout after Higher Secondary School =1 (baseline mean: 0.43)				
Specification 1: Combined treatment for partial and full exposure				
WB X Exposed	-0.094***	-0.088***	-0.095***	-0.072***
	(0.013)	(0.012)	(0.013)	(0.012)
Specification 2: Separate Dummies for partial and full exposure				
WB X Partially Exposed	-0.104***	-0.102***	-0.110***	-0.085***
	(0.015)	(0.015)	(0.011)	(0.012)
WB X Fully Exposed	-0.082***	-0.051***	-0.060***	-0.045***
	(0.016)	(0.014)	(0.016)	(0.013)
p-value: Test of Diff in Coeff	0.138	0.002	0.001	0.001
Observations	19,952	19,940	19,940	19,940
Controls	No	Yes	Yes	Yes
District FE	No	No	Yes	Yes
Birthyear FE	No	No	No	Yes
Exposed Cohort	Yes	Yes	Yes	No
West Bengal	Yes	Yes	No	No

Source: NSS Data. Bootstrapped standard errors are in parentheses, *** p<0.01, ** p<0.05, * p<0.1. We report the p-values from the test of the null hypothesis of equality between coefficients of partial exposure and full exposure.

Table 3: Heterogenous effects across gender and caste groups for the combined treatment

VARIABLES	(1) Technical	(2) STEM	(3) Dropout
Panel A: Heterogeneous effects across gender			
WB X Exposed X Fem	-0.070**	-0.077***	0.010
	(0.028)	(0.019)	(0.015)
WBXExposed	0.105***	0.082***	-0.076***
	(0.026)	(0.020)	(0.014)
female	-0.111***	-0.128***	0.037***
	(0.012)	(0.011)	(0.006)
Panel B: Heterogeneous effects across caste			
WB X Exposed X SC	-0.026	-0.010	-0.115***
	(0.026)	(0.025)	(0.013)
WB X Exposed X ST	-0.084	-0.134***	-0.115***
	(0.070)	(0.044)	(0.026)
WB X Exposed X OBC	0.018	-0.001	-0.092***
	(0.025)	(0.030)	(0.017)
WBXExposed	0.079***	0.056***	-0.027*
	(0.022)	(0.020)	(0.014)
SC	-0.026	-0.024	0.060***
	(0.016)	(0.017)	(0.009)
ST	-0.022	-0.012	0.036***
	(0.021)	(0.019)	(0.014)
OBC	-0.021	-0.017	0.035***
	(0.013)	(0.016)	(0.010)
Observations	8,545	8,545	19,940
Controls	Yes	Yes	Yes
District FE	Yes	Yes	Yes
Birthyear FE	Yes	Yes	Yes
Exposed Cohort	No	No	No
West Bengal	No	No	No

All estimates are for combined treatment in Phase 1 and Phase 2. The sample and all other specifications and covariates are the same as in Table 1 and Table 2. Bootstrapped standard errors are in parentheses.

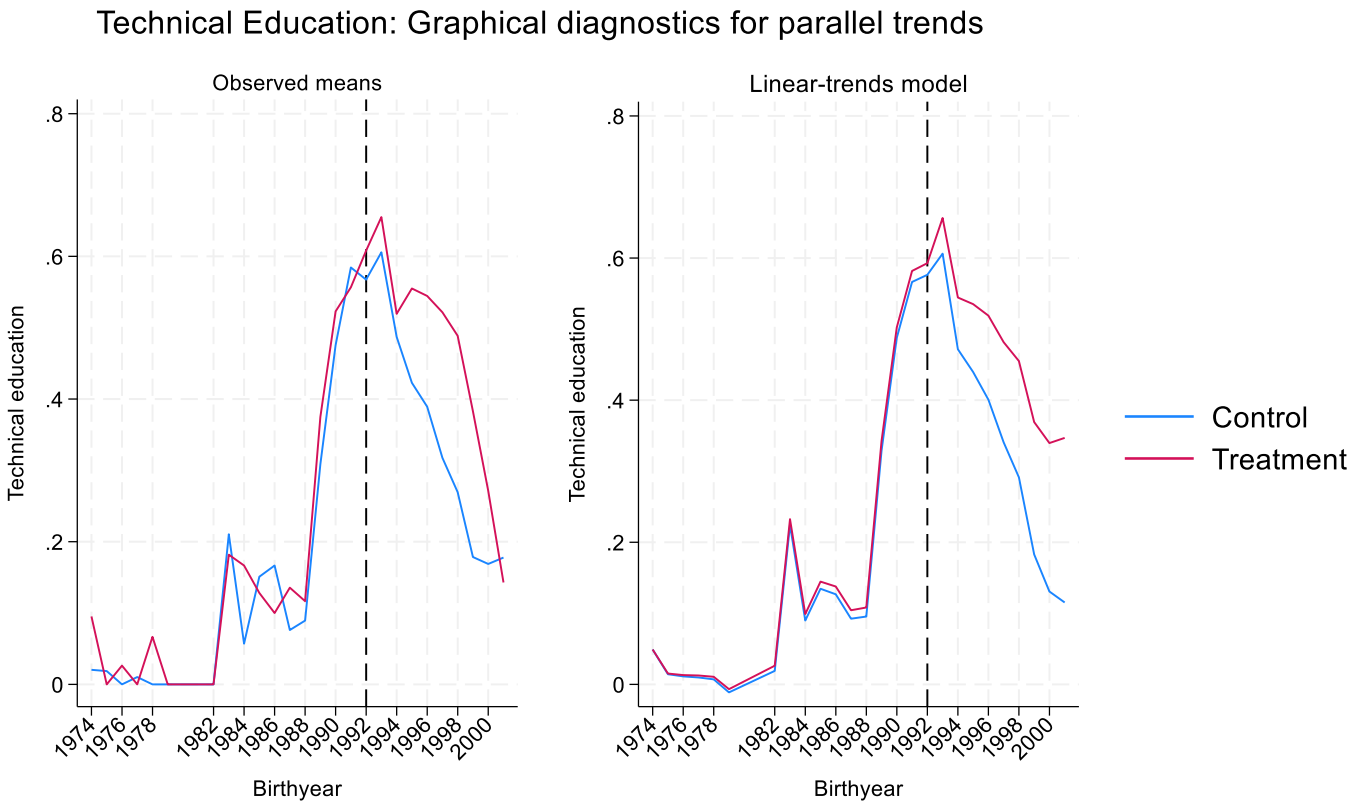
Table 4: Mechanism – Outcomes are English proficiency of individuals and indicator of writing skills in preferred language to capture the overall quality of education

	Mechanism	Potential confounder (Quality of Education)
	(1)	(2)
Outcomes	English proficiency	Overall writing skills in any preferred language
WB X Exposed	0.037**	-0.0000
	(0.015)	(0.0071)
Observations	14,388	12,438
R-squared	0.382	0.818
State FE	YES	YES
Birthyear	YES	YES

Bootstrapped SE in parentheses. *** p<0.01, ** p<0.05, * p<0.1
Source: IHDS 2 data(2011-12). Covariates Include State dummy, birth year dummy, Urban dummy, gender, caste and education of household head, Household Consumption Expenditure per capita, and Indicator if the mother tongue is the same as the state vernacular language.

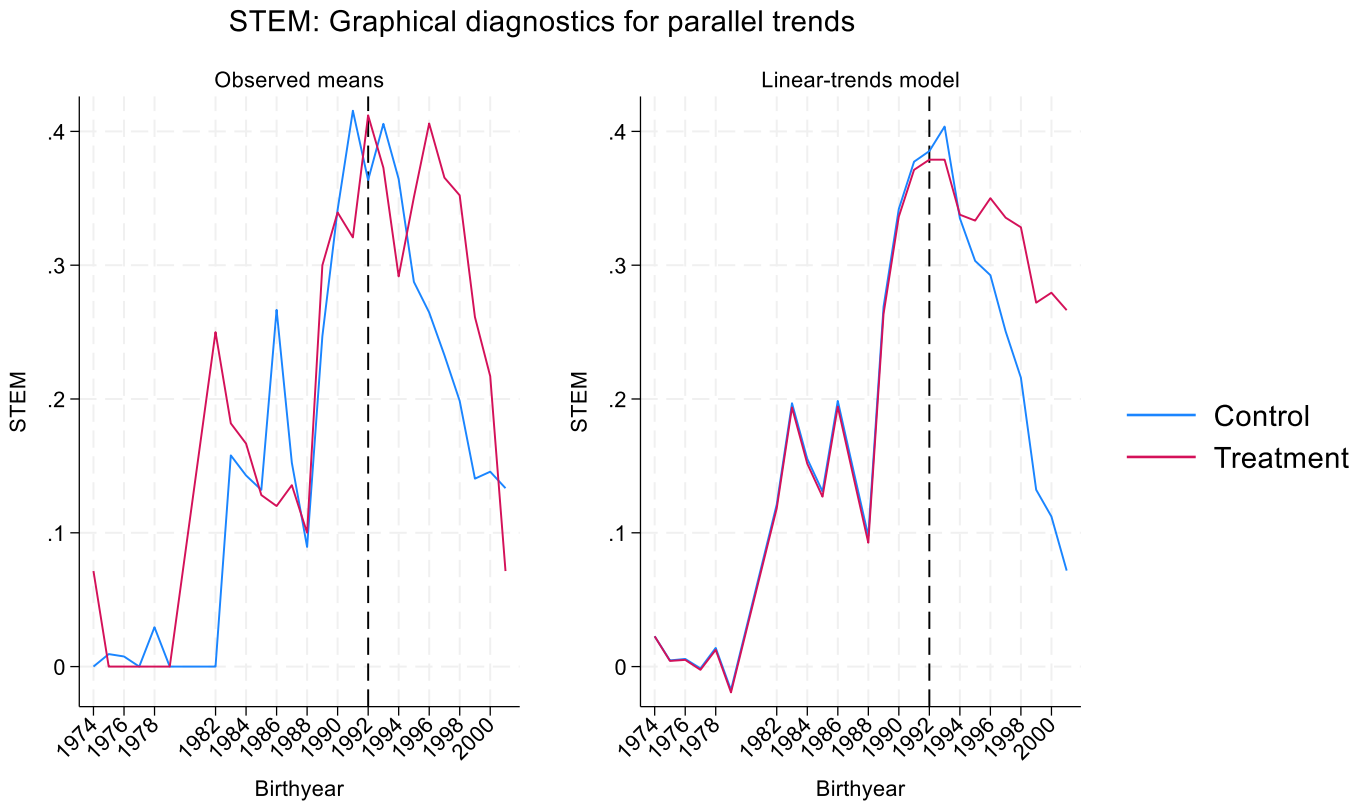
Using the same dataset, we test two hypotheses: 1. If improved English proficiency worked as a mechanism for choosing STEM by the exposed cohort. 2. Ruling out the overall change in the quality of education is the potential confounder of our treatment effects.

Figure 1A: Checking pretends by plotting observed means and fitting linear trends model for technical education



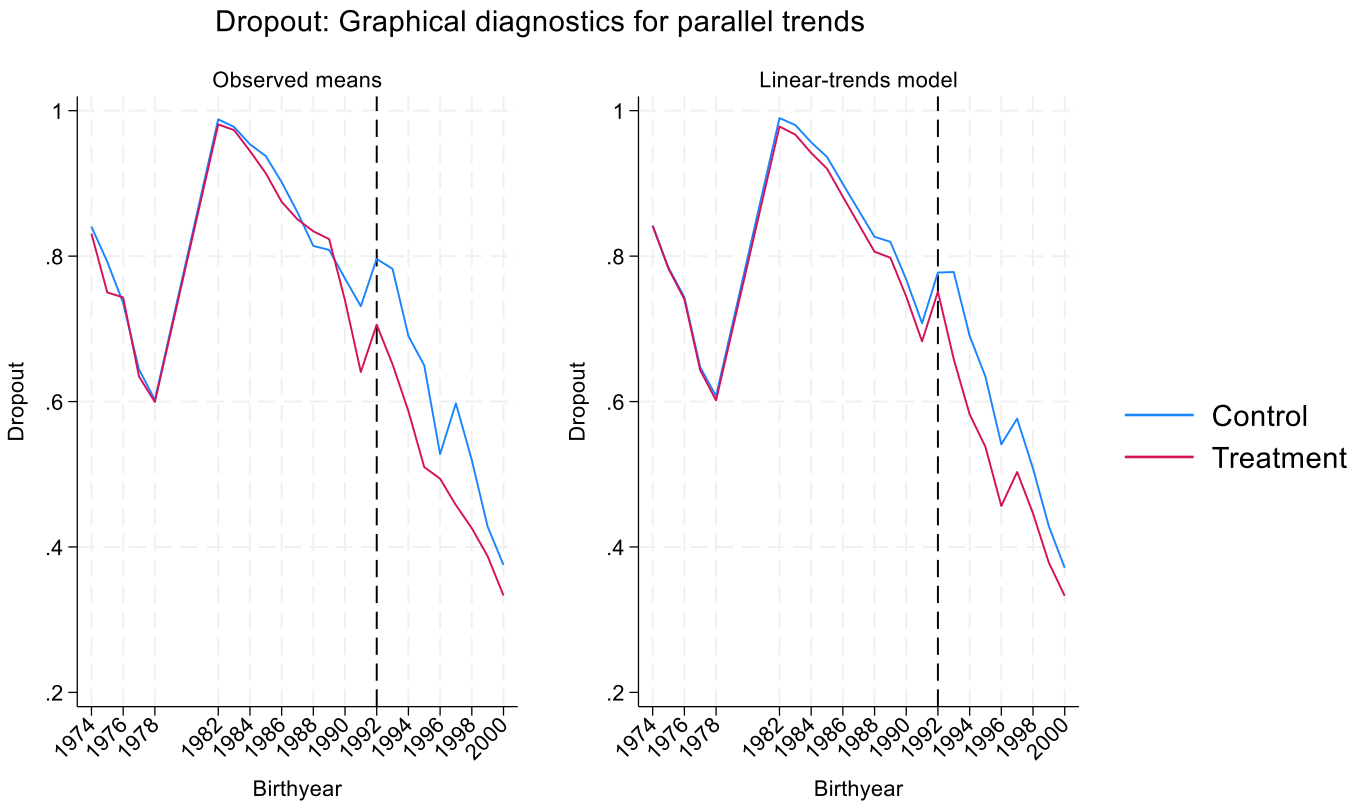
Data Source: NSS Rounds (Cross section) on Social Consumption known as Education Rounds as explained in the Data section. 1992 is the reference period. We control for the level effects only and do not control for any covariates here.

Figure 1B: Checking pretends by plotting observed means and fitting linear trends model for STEM education



Data Source: NSS Rounds (Cross section) on Social Consumption known as Education Rounds as explained in the Data section. 1992 is the reference period. We control for the level effects only and do not control for any covariates here.

Figure 1C: Checking pretends by plotting observed means and fitting linear trends model for dropout



Data Source: NSS Rounds (Cross section) on Social Consumption known as Education Rounds as explained in the Data section. 1992 is the reference period. We control for the level effects only and do not control for any covariates here.

Figure 2: Randomisation of treatment

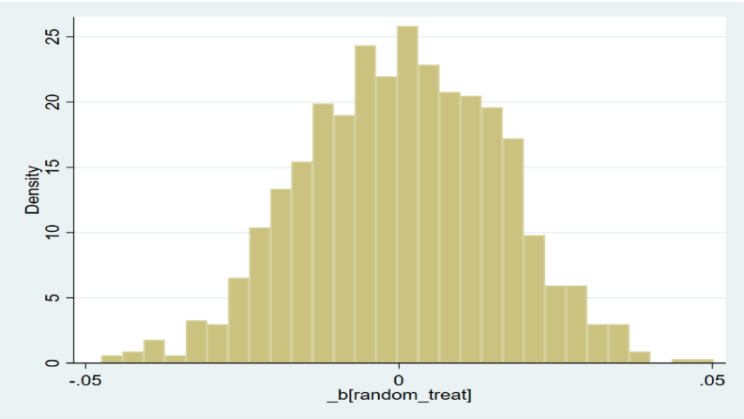


Figure 2a: Technical education

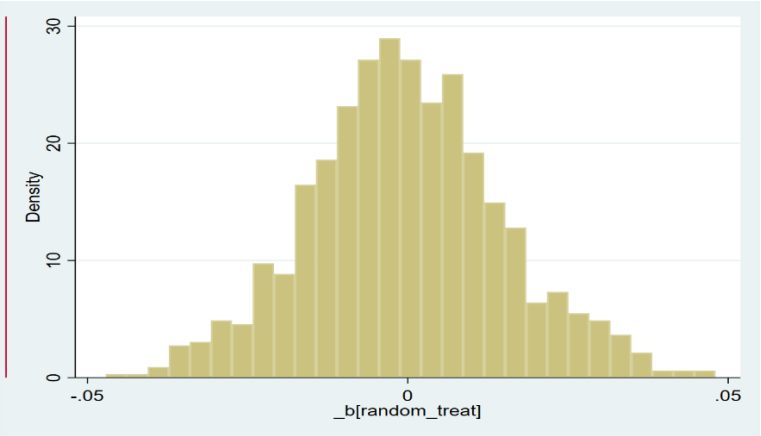


Figure 2b: STEM

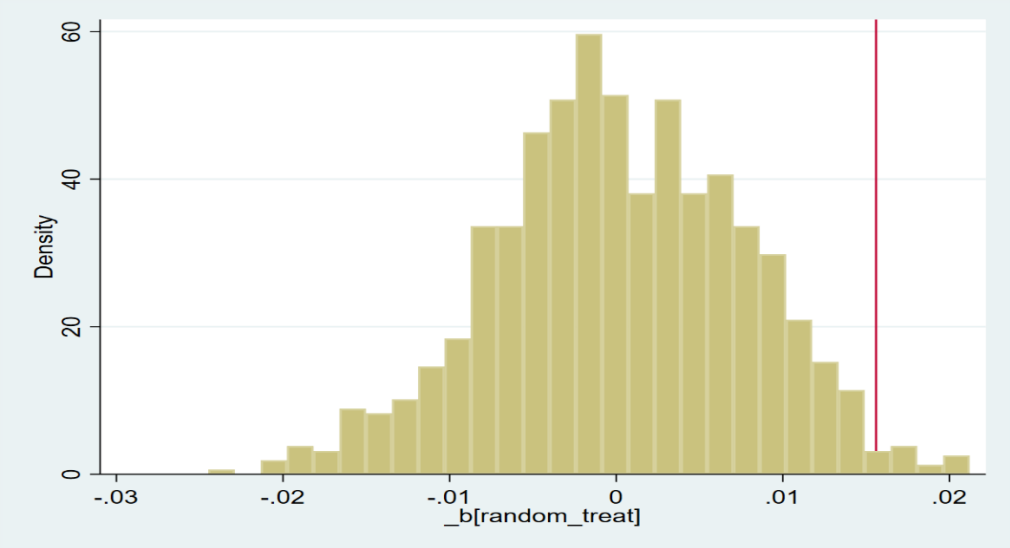


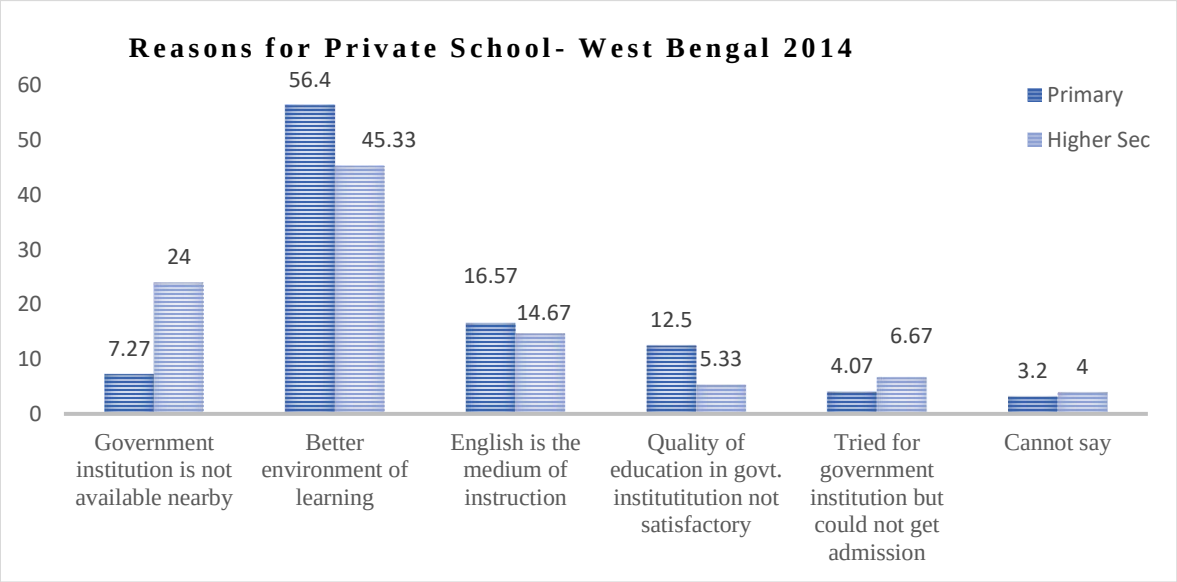
Figure 2c: Dropout

Note: The treatment is generated randomly 1000 times, and the treatment effects of the spurious treatment is plotted.

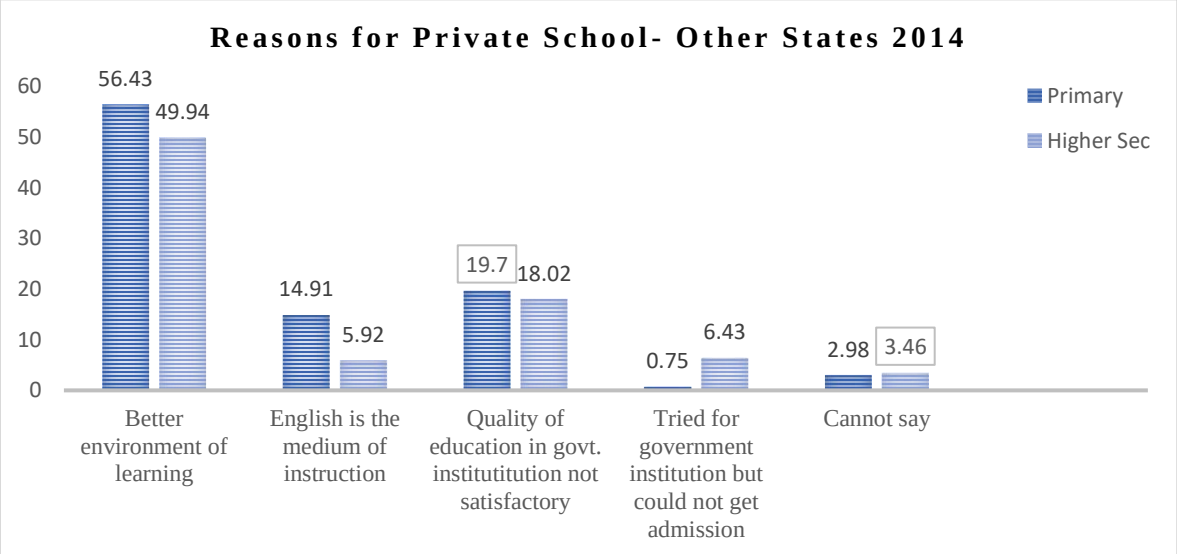
Appendix

Figure A1: Reasons for choosing private school by currently studying children in 2014

Sample of West Bengal

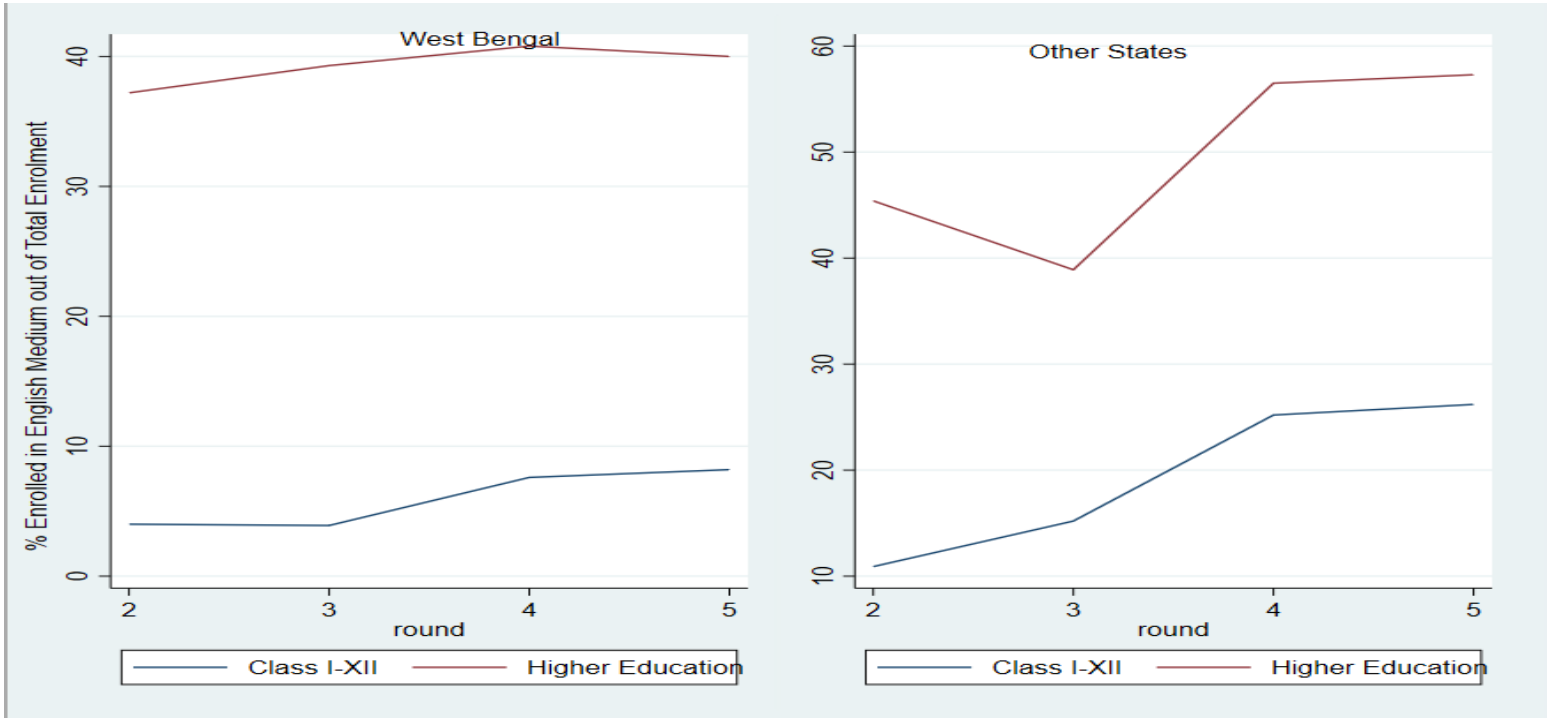


Sample of Other States



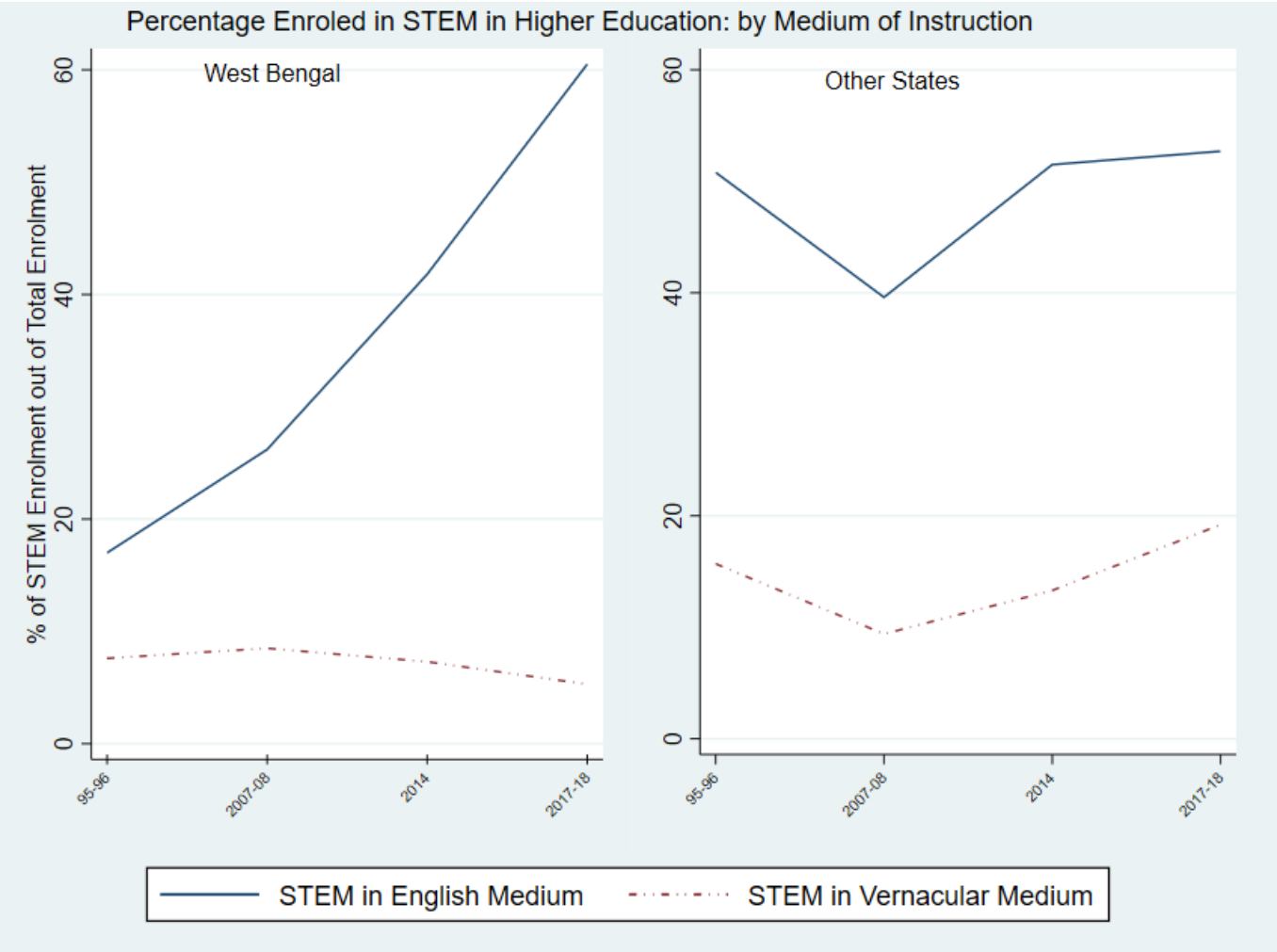
Notes: Data Source is 71st round of NSS conducted in the year 2014. Higher Sec / Primary: Currently Studying at Higher Secondary or Primary level. NSS does not ask questions on reasons for choosing private schools in earlier rounds of surveys.

Figure A2: Share of enrolment in English medium in West Bengal and Other states



Data Source: NSS Rounds (Cross section) on Social Consumption known as Education Rounds
The horizontal axis plots survey rounds 2, 3, 4, and 5, representing surveys conducted in 1995-96 (52nd round), 2007-08 (64th round), 2014 (71st round), and 2017-18 (75th round), respectively. The vertical axis plots percentage of total enrollment that was enrolled in English medium schools (Class 1-XII) or HE (Higher Education) institutions. The left panel is a sample of West Bengal, and right panel is a sample of other states.

Figure A3: Share of STEM enrolment in English medium versus vernacular medium HEIs



Data Source: NSS Rounds (Cross section) on Social Consumption known as Education Rounds

Figure A4: States of India with their Administrative Boundaries- Treated state (West Bengal) marked in dark shade and the Neighboring control states sarked in lighter shades



Table A1: Test of difference in means of outcomes

	Obs	Mean	SD	Obs	Mean	SD	Obs	Mean	SD	Difference	t-stat
Outcomes	All States			West Bengal (WB)			Neighboring States (NS)			(WB - NS)	
Technical	58,685	0.50	0.50	3,435	0.47	0.50	5,765	0.33	0.47	0.14	13.18
STEM	58,685	0.36	0.48	3,435	0.30	0.46	5,765	0.24	0.42	0.06	7.05
Dropout	127,675	0.32	0.47	7,630	0.33	0.47	15,821	0.36	0.48	-0.03	-5.29

Data Source: NSS rounds pooled together from 1995-96, 2007-8, 2014, 2017-18.

Table A2: Difference in Means of Covariates using pre-policy rounds of NSS

	West Bengal			**Control States**				
Covariates	Obs	Mean	SD	Obs	Mean	SD	Difference	t-stat
Female	581	0.49	0.50	1207	0.41	0.49	0.08**	3.15
Age	581	20.77	2.08	1207	20.31	2.04	0.46***	4.44
MPCE	581	1253.69	1093.0	1207	903.44	837.01	350.3***	7.48
Female head	581	0.52	0.50	1207	0.42	0.49	0.10***	3.85
Head secondary	581	0.58	0.49	1207	0.60	0.49	-0.02	-0.9
Distance	581	1.11	0.42	1207	1.19	0.55	-0.08**	-3.09
Rural	581	0.29	0.46	1207	0.42	0.49	-0.12***	-5.06
HH size	581	4.67	2.19	1207	5.82	3.27	-1.15***	-7.71
SC	581	0.15	0.36	1207	0.08	0.27	0.07***	4.88
ST	581	0.02	0.12	1207	0.07	0.25	-0.05***	-4.8

Note: NSS rounds pooled together from 1995-96, 2007-8

Table A3: Robustness to Gardner (2022)

VARIABLES	(1) Technical	(2) STEM	(6) Dropout
WBXExposed	0.082***	0.056*	-0.043**
	(0.025)	(0.030)	(0.021)
Observations	8,549	8,549	19,940
District FE	YES	YES	YES
Birthyear	FE	FE	FE

Sample: NSS pooled from all four rounds. Robust Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A4: Robustness to state-specific (time-variant or time-invariant) unobservables by dropping one control state at a time from the sample – Impact of English exposure on higher education outcomes

	(1)	(2)	(3)	(4)	(5)
Outcomes	Assam	Bihar	Jharkhand	Orissa	Sikkim
Technical	0.067***	0.069***	0.075***	0.085***	0.080***
	(0.021)	(0.021)	(0.019)	(0.024)	(0.020)
STEM	0.048***	0.020	0.055**	0.075***	0.052**
	(0.018)	(0.024)	(0.022)	(0.021)	(0.020)
Dropout	0.060***	0.076***	-0.053***	0.101***	0.071***
	(0.013)	(0.014)	(0.013)	(0.013)	(0.013)
Controls	Yes	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes	Yes
Birthyear FE	Yes	Yes	Yes	Yes	Yes
Exposed Cohort	No	No	No	No	No
West Bengal	No	No	No	No	No

Note: The column heads indicate the state that has been dropped from the sample. Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A5: Robustness to the removal of cohorts at margin-Impact of English exposure on HE outcomes with sample from birthyear 1977

VARIABLES	(1) Technical	(2) STEM	(3) Dropout
WBXExposed	0.083***	0.058**	-0.081***
	(0.022)	(0.024)	(0.012)
Observations	8,071	8,071	18,863
R-squared	0.292	0.206	0.357
Controls	Yes	Yes	Yes
District FE	Yes	Yes	Yes
Birthyear FE	Yes	Yes	Yes
Exposed Cohort	No	No	No
West Bengal	No	No	No

Note: This is a test of robustness using the full-model preferred specifications (the final columns of Tables 1 and 2). We reduce the sample by removing birth cohorts at the margin. That is, the birth years 1973-76 are removed from the samples of Tables 1 and 2.

*** p<0.01, ** p<0.05, * p<0.1