Monk Prayogshala Working
Paper # 2023-04

Bijoyetri Samaddar
Anirudh Tagat

August, 2023

LANGUAGE CAPITAL AND GRADE REPETITION: EVIDENCE FROM NATIONALLY-REPRESENTATIVE SURVEY DATA IN INDIA

monk prayogshala®

DEPARTMENT OF
SOCIOLOGY

Language capital and grade repetition: Evidence from nationally-representative survey data in India

Bijoyetri Samaddar
Department of Sociology, Monk Prayogshala, Mumbai, India

Anirudh Tagat
Department of Economics, Monk Prayogshala, Mumbai, India
School of Mathematics, Monash University, Melbourne, Australia

Address correspondence to Anirudh Tagat at at@monkprayogshala.in

The Sociology Working Papers are a series of ongoing research outputs from the Department of Sociology, Monk® Prayogshala®. The purpose of making these papers publicly available is to initiate dialogue and receive feedback on the preliminary work presented. The views and opinions expressed therein are those of the authors and do not necessarily reflect those of Monk Prayogshala, any supporting agencies, or any of its associated entities.

The authors declare no competing interests. Author Roles: BS - Conceptualization; Data curation; Investigation; Software; Roles/Writing - original draft. AT - Methodology, Software, Formal analysis, Supervision, Validation, Writing - original draft, Writing - review and editing

Language capital and grade repetition: Evidence from nationally-representative survey data in India

Abstract

This paper studies the impact of a discrepancy between the language spoken at home and the medium of instruction at school on educational outcomes using nationally representative sample surveys in India. Our sample consists of more than 100,000 students attending educational institutions. Using a fixed effects regression model, we find that language discrepancy is negatively associated with grade repetition among boys, and is not statistically significant for girls. Subgroup analyses suggest that results are driven by urban, relatively wealthier households residing in Hindi-speaking states. Implications for policies such as the recently proposed National Education Policy (NEP) are discussed.

Keywords: Native language instruction, education outcomes, National Education Policy (NEP), English language skills

Language capital and grade repetition: Evidence from nationally-representative survey data in India

1. Introduction

The Government of India rolled out the National Education Policy (NEP) in 2020, which lays significant emphasis on the parity between the medium of instruction in schools (until at least Grade 5) and the home language (or the mother tongue). This is seen as a policy push to improve student performance outcomes based on the assumption that schools should harness existing language capital in the classroom. However, there is little research on the links between language capital[1] and academic achievement in India, where there are 31 languages with more than 1 million native speakers as per the latest available data from the Census of India (Office of the Registrar General & Census Commissioner, India, 2011). Furthermore, data from a 2019-20 nationwide survey of schools revealed that nearly 26% of all children had English as the medium of instruction (Government of India, 2020). In such a diverse context, it is possible that learning in the native language could aid in enhancing learning outcomes. Such research is critical not only for policy but for investigating how human capital accumulation may vary by language.

In this study, we aim to establish a relationship between school performance and the discrepancy between home language and medium of instruction by using data from nationally representative sample surveys. Specifically, we use data from the National Sample Survey (NSS) 75th Round Schedule (2017-18) drawn from 36 states and union territories, consisting of 85,245 households, with 117,711 children enrolled from kindergarten to high school (Government of India, 2018). The 2011 Indian census defines mother tongue as the language spoken by the mother of a child or the

[1] In this paper, we use the terms language capital, linguistic capital, home language advantage to convey the same concept: that learning in a language that is spoken at home or that students are already proficient in can confer benefits in terms of knowledge and skill acquisition (Nag et al., 2018)

commonly spoken language at home during their childhood. The census also tells us that there are around 1369 rationalised languages in India (Khanuja et. al., 2021) According to the Eighth Schedule of the Indian Constitution, there are 22 scheduled languages[2]. There are demands for including another 38 languages[3] in the schedule. There is a common Indian saying that the language or language dialect spoken in an area changes every few hundred miles which highlights the extent of the linguistic variation in India. The use of Hindi, English and official state languages are mandated in the Constitution and the localized use of languages continues to happen in different regions. Thrasher (1996) writes that the British colonial rule in India resulted in the codification of many regional languages like Marathi and Bengali as a way of promoting "universal" languages, and thereby curbing, the linguistic variation. In postcolonial India, the early policy-makers intended Hindi to replace English as the official language of the sovereign nation. However, the states which did not commonly speak Hindi (like those in the southern or eastern parts of the country) resisted this decree. He adds that as a result of this, both Hindi and English emerged as the *lingua franca* in India. Hindi is commonly used in the northern states and state-wise regional languages are spoken in the rest of the country, whereas English is an overall common and official language of use.

There have been various policies enacted since the Independence of India that have attempted to both maintain linguistic diversity but at the same time find a common ground (or language if you will) for pedagogical and official functions. As early as 1968, through the National Policy on

[2] (1) Assamese, (2) Bengali, (3) Gujarati, (4) Hindi, (5) Kannada, (6) Kashmiri, (7) Konkani, (8) Malayalam, (9) Manipuri, (10) Marathi, (11) Nepali, (12) Oriya, (13) Punjabi, (14) Sanskrit, (15) Sindhi, (16) Tamil, (17) Telugu, (18) Urdu (19) Bodo, (20) Santhali, (21) Maithili and (22) Dogri.

[3] (1) Angika, (2) Banjara, (3) Bazika, (4) Bhojpuri, (5) Bhoti, (6) Bhotia, (7) Bundelkhandi (8) Chhattisgarhi, (9) Dhatki, (10) English, (11) Garhwali (Pahari), (12) Gondi, (13) Gujjar/Gujjari (14) Ho, (15) Kachachhi, (16) Kamtapuri, (17) Karbi, (18) Khasi, (19) Kodava (Coorg), (20) Kok Barak, (21) Kumaoni (Pahari), (22) Kurak, (23) Kurmali, (24) Lepcha, (25) Limbu, (26) Mizo (Lushai), (27) Magahi, (28) Mundari, (29) Nagpuri, (30) Nicobarese, (31) Pahari (Himachali), (32) Pali, (33) Rajasthani, (34) Sambalpuri/Kosali, (35) Shaurseni (Prakrit), (36) Siraiki, (37) Tenyidi and (38) Tulu.

Education in 1968, the Indian government had planned to follow a "three-language-formula" through which they foresaw Hindi eventually becoming the commonly spoken language across the nation. Though they also recognized the need to develop regional languages as a way of promoting cultural development across the country. This goal was reiterated in the 1986 draft of the education policy.

The National Education Policy (NEP) of 2020 upholds the "three-language formula" and emphasizes on the use of mother tongue language as the medium of instruction from grades 1 to 5. This has been met with a fair bit of criticism with some claiming that doing away with English language education in the early years of schooling would set children back and they would struggle to pick it up in later years and other seeing it as yet another move to enforce the Hindi as a pan-India language, especially as an instructional one (Jha, 2022). Saraswathy (2021) opines that most schools have teachers trained in English, Hindi or a regional language. Saraswathy adds that with this policy, it would mean that additional teachers would need to be trained and hired in order to make sure that there are enough teachers who can cater to the different mother tongues spoken commonly in an area which could vary greatly based on factors such as migration.

Research in different contexts have both shown that multilingual education has positive impact on children's learning outcomes and that English language skills has an important impact on earnings and opportunities (Azam et al., 2013; Ginsburgh & Weber, 2020). The extent of the various languages spoken in India, along with the vastly different socio-economic backgrounds of people serve as key determinants to the quality of education and learning outcomes that make it difficult to arrive at a one-size-fits-all solution as the NEP suggests. In this paper, we examine the congruence between the language spoken at home and the medium of instruction and its impacts on educational outcomes in India. This is in the backdrop of policies such as NEP that may not fully take into account the breadth and depth of cultural and linguistic differences, and how they might affect implementation as well. Data on educational achievement is not easily available alongside language data, and the dataset we use relies on a non-mainstream measure of educational outcome. Specifically, we use grade repetition as a proxy for school performance. Grade repetition or the process of being retained in a grade due to lack of academic performance is a common practice in many countries, including India. The consequences of grade repetition can be manifold. Ikeda and Garcia (2016) use data from OECD's Programme for International Student Assessment (PISA) to find that across all the countries from which data was collected, children who repeated a grade in secondary school were likely to have a lower reading performance as compared to those who did not repeat a grade. Kabay (2016) finds in her from her research on grade repetition and school performance in Uganda that repeating first or second grade does not have a significant impact on school dropout, but repeating third grade is significantly and positively related to dropping out of school. Kabay concludes that this might be the result of a Ugandan language policy, where children are taught in local languages until the third grade, and then from the fourth grade, the medium of instruction changes to English. This research highlights an important link between grade repetition and the transition of medium of instruction from a local language to English – something that India's NEP 2020 also plans to implement.

The remainder of this paper is organised as follows. Section 2 contains the review of literature in this area and more background on the Indian education system. Section 3 provides an overview of the data, and describes some of the summary statistics. Section 4 provides the empirical framework within which the research questions are addressed. Section 5 showcases the results and discusses them in the context of existing work. Section 6 concludes and provides implications for policy and directions for future research.

2. Literature and Background

Much research has been conducted on the concordance between mother tongue and language of instruction especially in the context of African nations, which have many linguistically and ethnically diverse societies. For example, in Ethiopia, Agraw (2016) found that the introduction of mother tongue-based education in 1994 led to significant improvements in reading skills and early labour market outcomes for birth cohorts that gained access to it. The provision of primary education in mother tongue halved the reading skills gap between Amharic and non-Amharic mother tongue users in Ethiopia. Seid (2017) finds that children taught in their mother tongue in grades 1 to 4 in Ethiopia have better mathematics test scores than the ones who are not taught in their mother tongue after they transition to an English medium of instruction in grade 5. In South Africa, Eriksson (2014) found that a change in the language of instruction in schools in 1955 from English or Afrikaans to mother-tongue instruction had positive effects on wages, educational attainment, and the ability to read and write. In Cameroon, Laitin et. al. (2019) found that instructing children in the local language during the first three years of school showed positive gains in English and Math skills as well as an increase in probability of staying on in school after the time period of the intervention. Bernhofer and Tonin (2022) use data from a unique educational set up in Germany to show that taking an exam in a language other than their native tongue (which could be Italian, German, or English) leads to a sharp decline in performance. Many studies in economics look at the impact of an official change in language of education policy (J. Angrist et al., 2008; J. D. Angrist & Lavy, 1997) in countries.[4] These results are somewhat mixed, in that they are context-specific and suggest

[4] For a review of economics work in the domain of language, we refer the reader to Ginsburgh and Weber (2020), which outlines many areas of theoretical and empirical work looking at the role of language.

that there is no overwhelming evidence that home language instruction or English instruction confers a significant advantage in the short-term.

While the exploration of how language of instruction impacts school performance in the Indian context remains, there has been some key research conducted to identify the consequences of language mismatch in education and its consequences on school achievement. Jain (2017) uses waves of census data in India to identify how mother tongue impacts educational attainment in South India before and after the states were reorganized (pre- and post-Independence of India) along the lines of linguistic and cultural similarities. He finds that the colonial structuring of states (which were conducted in a rather disorganised manner) where linguistic minority groups were mixed with linguistic majority groups in the same state led to a difference in educational attainment which was to some extent remedied by reorganization of states in postcolonial South India.

In addition to the diversity of languages spoken in India, there is also a deep-seated belief that gaining English language skills is an important way to gain social mobility and economic returns. Kapur and Chakraborty (2008) find that the revoking of English language education in government schools in West Bengal resulted in lower wages among those who were impacted by the policy. Along a similar vein, Azam et. al. (2013) find that there is a strong positive relationship between English-language skills and earnings after controlling for factors such as age, schooling, social group, etc., where men who are fluent English speakers earn much higher than those who speak no English. Men who are not as fluent in English also earn higher than those who speak no English. Tsimpli et. al. (2020) highlight the importance of linguistic diversity in cognitive skills through their findings where they note that bilingual children perform better in cognitive tasks as compared to monolingual children and monolingual children who are exposed to linguistic diversity in their environment have positive impact on their cognitive performance.

Finally, systematic evidence on this area of work by Nag et al. (2018) shows considerable work in favour of the home language advantage hypothesis. They discuss the role of the Home Language and Literacy Environment (HLLE) as a key determinant of whether there is a home-language advantage at school. However, a majority of the studies that they review are drawn from high-income countries, whereas in ethnically diverse, low-income countries, they find mixed evidence about the impact of such interventions (that involve providing textbooks, or home tutoring). Of the studies included in this review, only a few look at India, and are largely reliant on qualitative methods that help understand the role of specific home-based factors in improving educational outcomes for children. Our study aims to contribute to the literature on empirical evidence on the language discrepancy and impacts on educational outcomes in the Indian context. We build on current work by using a nationally-representative large-scale household survey dataset from 2017-18 and using a robust fixed effects regression analyses.

3. Data

The data used for this study are taken from the Household Social Consumption on Education in India, which is collected through the National Sample Surveys (NSS) by the Government of India. The datasets are specifically from the 75th round of surveys conducted between July 2017 to June 2018. The survey has a combination of both qualitative and quantitative questions in order to ascertain the educational attainment of household members. Some of the qualitative questions include the type of institution attended by participants, their current education level, the type of transport they use to travel, whether they receive midday meals, and the type of labour their household was engaged in. The quantitative questions include total expenditure costs on education, scholarship amounts received (if any), and the household's usual monthly consumer expenditure. The survey covers rural and urban areas from all districts within all Indian states and union territories, except for some of the remote villages and areas in the Andaman and Nicobar Islands, which is a union territory of India that lies on the Bay of Bengal.

The survey uses a stratified multi-stage design. The first stage units (FSUs) were primarily the census villages (*panchayat* wards in the case of Kerala) in rural areas and Urban Frame Survey (UFS) blocks in urban areas. If the FSU was large, then another step was undertaken to select two smaller sub-groups known as hamlet-groups (hgs) in rural areas or sub-blocks (sbs) in urban areas. The next stage of units, also called the ultimate stage units (USUs), were the households in rural and urban areas. For more details on the sampling methodology, we refer the reader to the NSSO documentation.

For preparing our main dataset for analysis, we used the different datasets from each of the blocks. Blocks 4, 5 and 6 have different information components of the survey collected. These blocks have information at the individual level – demographic particulars of participants, education particulars of participants aged between 3 and 35 years who are attending school currently and expenditure particulars of those attending school between the ages of 3 and 35 who are attending school (pre-primary and above) respectively. These were done as they would be key for our analysis. Block 3 has information at the household level – more specifically variables on household characteristics such as the religion or social group of the household, the household size, the source of income of the household, etc. In order to merge these datasets, we used the household ID variable ("HHID") and serial number ID of each participant in the household ("Per_serialno") to merge the individual-level datasets (of blocks 4, 5 and 6). We used just the household ID variable ("HHID") to merge a consolidated dataset (containing the individual-level characteristics) with the block 3 dataset containing the household characteristics.

In its consolidated form, the dataset contains 513,366 of all participants who were interviewed for the survey. Out of this, about 152,992 participants between the ages of 3 and 35 are currently attending any form of educational training. The documentation pertaining to the survey states that any household member would be considered as student if they fall between the ages of 3 to 35 years and are currently attending education. The variable "Course_attending" from block 5 gives the distribution of the sample on the basis of whether they are currently enrolled into school upto class X (grade 10) or are currently enrolled in "Science", "Humanities" or "Commerce". Since the Indian education system allows students to pick streams of study after grade 10 which can be classified into the three aforementioned categories, we also consider these students as part of the sample. Almost 93,444 (61.25 per cent) of the school-going sample are enrolled in a grade upto 10. Those in the three streams of study comprise 28,527 participants (nearly 19 per cent of the school-going sample). Table 1 contains the description of the key sample characteristics.

Table 1: Summary statistics

	Mean	SD	N
Age	30.08	19.25	513365
Female	0.48	0.50	513366
Household size	5.30	2.37	513366
Number of children	1.50	1.40	513366
Proportion attending school (up to 10th grade)	0.24	0.43	513366
Proportion repeating same grade	0.03	0.16	150250
Language discrepancy	0.09	0.29	513366
Linguistic distance	0.34	1.46	77865
Average education level of Father (years)	8.06	4.34	513342
Average education level of Mother (years)	6.37	4.65	513342
Average education level in household (years)	6.64	2.82	513342

Transport costs (INR)	1949.55	3352.85	97545
Private coaching costs (INR)	2170.23	5410.24	64245
Books and stationary cost (INR)	1812.22	2036.02	150583
Course and fees cost (INR)	6932.66	17041.82	130050
Other school-related expenses (INR)	649.34	2101.30	122472
Monthly household consumption expenditure (INR)	10471.98	7370.85	513366
School characteristics			
Proportion attending government school	0.59	0.49	152558
Proportion attending private-public school	0.13	0.34	152558
Proportion attending private school	0.28	0.45	152558
Distance < 3kms to school	0.19	0.39	513366
Distance between 3-5kms to school	0.02	0.12	513366
Distance >5kms to school	0.04	0.21	513366
Proportion receiving scholarship	0.14	0.34	152558
Proportion taking private tutoring	0.20	0.40	152558
Religion			
Hindu	0.81	0.39	513352
Muslim	0.14	0.34	513352
Christian	0.02	0.15	513352
Sikh	0.02	0.13	513352
Jain	0.00	0.05	513352
Buddhist	0.00	0.07	513352
Parsi	0.00	0.01	513352
Other	0.00	0.06	513352
Social Grouping			
Scheduled Caste	0.10	0.29	513366
Scheduled Tribe	0.19	0.40	513366
Other Backward Class	0.45	0.50	513366
Upper Caste / General	0.26	0.44	513366

Source: NSS Household Schedule 25.2: Social Consumption: Education
Note: Sample weights applied.

In the appendix, Table A.1 reports the descriptive statistics around language spoken at home as well as the medium of instruction. We note that our sample has a majority of Hindi-speaking households (46% approximately), and virtually no English-speaking households. In contrast, 27% of households report sending their children to a school where English is the primary medium of instruction, which is still lower than the majority of Hindi-medium school attending children (44% of the sample). The next highest languages in terms of spoken at home as well as medium of instruction are Bengali (8% of the sample speak at home) and Marathi (7% of the sample speak at home).

We next outline key educational and socio-demographic statistics that could vary by the language discrepancy. We do an unweighted t-test of these indicators by whether or not a child faces a language discrepancy. Note that the statistical significance of these tests cannot be interpreted as per other t-tests. Table 2 contains the results of this exercise. Since these are unweighted t-tests, the statistical significance cannot be interpreted in a straightforward manner. When we repeated the same tests with weighted regressions on each variable, however, we found that the proportion of our sample repeating the same grade, (p = 0.195) and those speaking English at home (p = 0.17), and proportion of households belonging to the scheduled caste (p = 0.01) were not statistically significant after correcting for multiple hypothesis testing (threshold p < between the two groups.

Table 2: Difference between language discrepancy and concordance groups

	N_0	$Mean_0$	N_1	$Mean_1$	t-stat
Age	434476	33.46	78889	14.60	541.75
Female	434477	0.49	78889	0.41	39.62
Household size	434477	5.46	78889	5.19	31.33
Number of children	434477	1.45	78889	1.09	75.95
Proportion attending school (up to 10th grade)	434477	0.15	78889	0.71	-322.83
Proportion repeating same grade	73143	0.03	77107	0.03	-4.30
Average education level of Father (years)	434459	8.53	78883	9.71	-72.93
Average education level of Mother (years)	434459	6.85	78883	7.69	-45.31
Average education level in household (years)	434459	7.30	78883	8.56	-123.41
Transport costs (INR)	42168	1510.80	55377	4097.03	-100.87
Private coaching costs (INR)	34371	1834.08	29874	3603.41	-31.17
Books and stationary cost (INR)	72558	1564.61	78025	3305.02	-120.16
Course and fees cost (INR)	58094	5134.93	71956	19672.51	-106.24
Other school-related expenses (INR)	58929	559.95	63543	1473.34	-51.87
Monthly household consumption expenditure (INR)	434477	11890.03	78889	14800.67	-75.80
School characteristics					
Proportion attending government school	73815	0.71	78743	0.38	135.38
Proportion attending private-public school	73815	0.12	78743	0.19	-38.96

Proportion attending private school	73815	0.17	78743	0.43	-112.68
Distance < 3kms to school	434477	0.13	78889	0.55	-227.38
Distance between 3-5kms to school	434477	0.01	78889	0.08	-69.17
Distance >5kms to school	434477	0.03	78889	0.37	-192.50
Proportion receiving scholarship	73815	0.17	78743	0.14	21.02
Proportion taking private tutoring	73815	0.19	78743	0.17	11.07
Language spoken at home					
Hindi	74010	0.54	78873	0.23	131.59
English	74010	0.00	78873	0.00	7.23
Gujarati	74010	0.05	78873	0.02	33.75
Kannada	74010	0.03	78873	0.03	-6.27
Marathi	74010	0.07	78873	0.05	13.80
Tamil	74010	0.04	78873	0.07	-25.34
Telegu	74010	0.03	78873	0.08	-40.88
Medium of instruction					
Hindi	74010	0.54	78813	0.10	206.04
English	74010	0.00	78813	0.83	-618.73
Gujarati	74010	0.05	78813	0.00	57.03
Kannada	74010	0.03	78813	0.01	32.49
Marathi	74010	0.07	78813	0.01	63.97
Tamil	74010	0.04	78813	0.00	52.11
Telegu	74010	0.03	78813	0.00	43.02
Religion					
Hindu	434468	0.76	78884	0.68	45.75
Muslim	434468	0.14	78884	0.14	-0.01*
Christian	434468	0.06	78884	0.12	-48.84
Sikh	434468	0.02	78884	0.03	-9.86
Jain	434468	0.00	78884	0.00	-6.15
Buddhist	434468	0.01	78884	0.01	-12.14
Other	434468	0.01	78884	0.02	-20.39
Social Grouping					
Scheduled Caste	434477	0.14	78889	0.18	-29.14
Scheduled Tribe	434477	0.17	78889	0.11	45.67
Other Backward Class	434477	0.41	78889	0.36	23.09
Upper Caste / General	434477	0.28	78889	0.34	-32.14

Source: NSS Household Schedule 25.2: Social Consumption: Education

Note: N_0 refers to the number of observations for the group that does not have any language discrepancy, whereas N_1 refers to the number of observations for the group with language discrepancy. t-test conducted without sample weights and for samples with unequal variances. All t-tests were statistically significant at the 1% level, except proportion of Muslim households, which was not statistically significant.

4. Methodology

Ideally, we could examine the causal impact of the concordance between mother tongue and language of instruction on education outcomes if we could exploit some natural variation in language of instruction via a policy or similar. Another way to understand causal impacts involves randomly assigning some students to receive instruction in their mother tongue and other students to receive instruction in another language. Neither of these are available or feasible in our context, and we thus rely on a fixed effects regression approach to examine the relationship between language discrepancy and educational outcomes using secondary data. We note that these are only interpretable as associations or correlations, and cannot be modelled as causal impacts. However, the fixed effects approach helps control for a range of other factors that could also influence grade repetition, helping to reasonably isolate the association with language discrepancy. As prior work in economics and education shows, this is not straightforward, especially since there are likely to be a range of confounding factors that could affect educational outcomes as well as school choice (which determines medium of instruction). Thus, since choice of educational institution (and therefore the language of instruction) is self-selected, we cannot provide causal estimates of the impact of the disparity between mother tongue and the language of instruction on educational outcomes. We preface our empirical framework by acknowledging that disentangling causality from this framework is challenging. However, there is value in exploring associations between language capital and educational outcomes employing a regression estimation using data outlined previously.

We estimate an education production function, where the main outcome of interest (and the only variable that proxies for educational achievement) is whether the student is repeating the same grade as they were in last year. Research typically uses a range of individual, household, and school characteristics to explain educational outcomes. Where data on academic achievement scores (or grades) would be available, we could have used an ordinary least squares (OLS) estimator, but since our dependent variable is binary, we instead use a linear probability model (LPM) and use logistic and probit regressions to check for robustness of the functional form. Specifically, we estimate the equation:

$$Repeat_{ihs} = \alpha + \beta_1\, Language_{ih} + \beta_2\, School_{is} + \beta_3\, Child_{ihs} + \beta_4\, HH_h + \epsilon \qquad (1)$$

Where,

$Repeat_{ihs}$ is a binary variable that takes a value of 1 if the ith child in the hth household residing in the

specific characteristics, such as the type of institution (government −

owned or privately owned), the distance to school, and the mode of transport used to attend school. Sir

day meal at school, whether they received any government benefits (e.g., free textbooks or a scholarsh

includes variables on parental education qualifications, religion/caste, location (urban or rural), and an indicator of household wealth (usually measured as the monthly per capita consumption expenditures). The standard errors are clustered at the level of the household to account for any correlation among unobservables within the household.

The main coefficient of interest is β_1, which will tell us about the extent to which linguistic capital (measured as the concordance between mother tongue of the child and medium of instruction at school) is associated with the likelihood of repeating a grade in India. We should note that there are various other factors that could influence our outcome variable on which we do not have data. For example, there could be household-level shocks, or other exogenous shocks (such as droughts, floods, or extremely high temperatures) that could also affect educational outcomes. We do not have data on shocks, and instead include household characteristics that could capture some of these impacts, as well as state fixed effects for any state-level exogenous shocks or changes in policy. The likelihood of grade repetition could also be tied to health status and indicators, on which we do not have corresponding data for the child or other members of the household. Lastly, the availability of work at the time of schooling could contribute to educational outcomes as well, and we include some data on whether the child was looking for work when attending school to account for this in an additional specification.

We run these analyses separately for rural and urban residents, level of schooling, and also for quintiles of household consumption expenditures (as the closest proxy for household income and wealth). Finally, to check whether this disparity is more concentrated where English is the language of instruction, we run estimation of equation (1) separately for children attending English and non-English medium of instruction schools, as well as whether Hindi is the dominant language spoken in that state or otherwise.

5. Results

We first describe results from the estimation of equation (1) using LPM in Table 3, without fixed effects and without additional educational inputs (parental education and private tutoring) in Columns 1-2 for boys and girls separately. Next, columns 3-4 we include these additional educational inputs. Columns 5-6 report the same estimates as 3-4, but with state fixed effects. Columns 7-8 report the estimation with added language spoken at home fixed effects, and finally, Columns 9-10 report results with medium of instruction at school fixed effects.

Table 3: Regression results of language discrepancy on grade repetition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Language discrepancy	-0.001	0.005	-0.004	0.005	-0.007**	0.001	-0.004	0.003	-0.012**	0.006
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.006)	(0.010)
ln (Father's education level)			0.001	-0.002	0.001	-0.002	0.001	-0.002	0.001	-0.002
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
ln (Mother's education level)			0.001	0.001	-0.000	-0.001	-0.000	-0.001	-0.000	-0.001
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Government school	-0.005	-0.011***	-0.007	-0.015***	-0.014***	-0.023***	-0.013**	-0.022***	-0.012**	-0.022***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
Distance to school	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.001	0.000	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
ln (Age)	0.012***	0.018***	0.013***	0.019***	0.014***	0.021***	0.014***	0.022***	0.015***	0.022***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Scholarship recipient	-0.003	-0.008**	-0.004	-0.011***	-0.007	-0.014***	-0.006	-0.013***	-0.006	-0.013***
	(0.005)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)
Takes private coaching			-0.003	0.005	-0.004	0.003	-0.003	0.003	-0.003	0.003
			(0.003)	(0.004)	(0.003)	(0.005)	(0.003)	(0.005)	(0.004)	(0.005)
ln (Total expenditure on education)	-0.002	-0.007***	-0.002	-0.008***	-0.003	-0.009***	-0.003	-0.009***	-0.004*	-0.009***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)
ln (Household size)	-0.006	-0.004	-0.000	-0.007	0.002	-0.004	0.002	-0.005	0.002	-0.005
	(0.006)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
ln (Monthly household consumption expenditure)	-0.001	-0.002	-0.000	-0.001	0.000	0.001	-0.000	-0.000	-0.000	0.000
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)	(0.005)
Constant	0.029	0.059**	0.022	0.081**	0.021	0.073**	0.025	0.079**	0.030	0.076**
	(0.031)	(0.029)	(0.034)	(0.032)	(0.037)	(0.035)	(0.038)	(0.035)	(0.039)	(0.036)
State Fixed Effects	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
Language Fixed Effects	No	No	No	No	No	No	Yes	Yes	Yes	Yes
Medium of Instruction Fixed Effects	No	No	No	No	No	No	No	No	Yes	Yes

Observations	85,390	63,317	78,638	58,387	78,638	58,387	78,566	58,356	78,530	58,325
R-squared	0.003	0.004	0.003	0.005	0.008	0.011	0.011	0.015	0.012	0.016

Note: Each column contains coefficients of the independent variable in that row on the main dependent variable, grade repetition (which is a binary variable, taking value of 1 if the child was repeating the same grade as last year). All models are estimated using the linear probability model, and include additional controls for religion and social group, whose coefficients are not reported here. Sample weights are applied. Standard errors clustered at the level of the household reported in parentheses; *** p<0.01, ** p<0.05, * p<0.1

First, we note that language discrepancy is not statistically significantly associated with grade repetition for girls across all specifications. Although the coefficient is positive, our results suggest that there is no detrimental impact of language discrepancy for girls. Notably, we find a positive and statistically significant association between language discrepancy for boys in the specifications that include state fixed effects, as well as the full specification that has state, medium of instruction, and language spoken at home fixed effects. Here, we find that when the language spoken at home is not the same as the medium of instruction at school, there is a 1.2 percentage point *lower* likelihood of grade repetition (Column 9 in Table 3). This result is statistically significant at the 5% level. This suggests that language discrepancy is not detrimental to academic performance at school, but more so that it is likely to signal a lower chance of grade repetition. This could be on account of various reasons, which we will explore in the next subsection on mechanisms.

Other results show that being in a government institution (relative to those in private or public-private schools) is associated with a lower likelihood of grade repetition, for both boys and girls. There is a 1.2 percentage point reduction (2.2) for boys (girls) in the likelihood of grade repetition for children attending a government school. This is likely linked to grading policies or grade retention in government schools in the survey period. Since our measure of school outcome is related to grade repetition (which does not reflect outcomes at a point in time, but typically over a period of time), the distance to school is not significantly associated with grade repetition. In contrast, being a scholarship recipient (which are sometimes especially awarded to girl children to encourage them to stay in school), is significantly associated with a lower likelihood of grade repetition among girls.[5] There is no similar association observed for boys. In terms of other

[5] This is also suggesting that scholarships themselves may be less likely to be awarded to girls who are repeating grades, but since we are working with cross-sectional data, we have no way to verify this.

educational inputs, we look at whether the student takes private tutoring or coaching to assist with their schooling, and the total educational expenditure incurred by the household. The former is not statistically significant in the model across all specifications, suggesting that private coaching is not associated with grade repetition in our sample. In contrast, educational expenditures are statistically significantly associated with a lower likelihood of grade repetition among both boys and girls. In the full specification, a 1 percentage point increase in education expenditures reduces the likelihood of grade repetition by 0.4 percentage points for boys and 0.9 percentage points for girls. This result is broadly in line with various other studies on the role that household investment in education plays, except that ours is perhaps the first to directly link expenditures with grade repetition in the Indian context.

5.1 Potential Mechanisms

We now turn to a rich set of subgroup analyses to help infer the mechanisms through which language discrepancy could be associated with grade repetition in our sample for boys and girls. We are especially interested in exploring what is driving the lower likelihood of grade repetition when there is a language discrepancy for boys, as highlighted in the previous section.

Table 4a contains the set of heterogeneous analyses, where the model that is estimated is the full model, including state, medium of instruction, and language spoken at home fixed effects. While all covariates as in equation (1) are included, we only report the coefficient on language discrepancy to streamline the discussion. The first two columns report the results for boys and girls by sector of residence (urban and rural). We find that the negative association between language discrepancy and grade repetition (i.e., language discrepancy is less associated with grade repetition among boys) is dominated by households residing in urban areas. The corresponding coefficient is 0.019, and is statistically significant at the 10% level. This suggests that where the medium of instruction in schools differs from the mother tongue in urban areas, there is less likelihood of grade repetition for boys only. The corresponding result for girls remains consistent with the main results, suggesting that this association is stronger for boys in urban households.

Next, to examine whether household wealth may be driving some of these results, we split the sample into quartiles of monthly consumption expenditure (Columns 5-12 in Table 4a). The first quartile (expenditures averaging INR 5142 or USD 80 approximately) per month, as well as the second (expenditures averaging INR 8793 or USD 136 approximately), and third (expenditures averaging INR 12760 or USD 200 approximately) have no statistically significant coefficients. However, in the households in the top 25% of consumption expenditures (averaging INR 24039 or USD 375 approximately), the statistically significant and negative association between language discrepancy and grade repetition for boys is found. Among the households with high consumption expenditure (i.e., potentially the richest households), language discrepancy is associated with a 3.8 percentage point reduction in the likelihood of grade repetition. There is no corresponding association found for girls in these households.

Table 4a: Subgroup analysis by sector of residence and consumption quartiles

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Rural		Urban		1st Quartile		2nd Quartile		3rd Quartile		4th Quartile	
VARIABLES	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Language discrepancy	-0.007	0.009	-0.019*	-0.013	-0.002	0.006	0.001	0.011	-0.011	0.002	-0.038**	0.004
	(0.009)	(0.016)	(0.010)	(0.010)	(0.014)	(0.023)	(0.011)	(0.017)	(0.008)	(0.010)	(0.017)	(0.028)
Constant	0.131**	0.149***	-0.034	0.053	0.381**	0.332***	-0.337	-0.073	-0.207	0.128	0.012	0.048
	(0.060)	(0.050)	(0.048)	(0.060)	(0.165)	(0.128)	(0.226)	(0.227)	(0.256)	(0.221)	(0.098)	(0.112)
Observations	47,052	34,387	31,477	23,937	18,943	13,153	20,965	15,344	18,624	14,199	19,992	15,621
R-squared	0.013	0.027	0.033	0.022	0.028	0.043	0.019	0.022	0.019	0.024	0.067	0.043

Note: Each column contains coefficients of the independent variable in that row on the main dependent variable, grade repetition (which is a binary variable, taking value of 1 if the child was repeating the same grade as last year). All models are estimated using the linear probability model, and include additional controls for religion and social group, whose coefficients are not reported here. Sample weights are applied and state, medium of instruction, and language spoken at home fixed effects are included. Standard errors clustered at the level of the household reported in parentheses; *** p<0.01, ** p<0.05, * p<0.1.

Table 4b: Subgroup analysis by medium of instruction, Hindi language in state, and schooling level

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Non-English medium schools		Hindi-speaking states		Non-Hindi speaking states		Up to 10th grade		12th grade and above	
VARIABLES	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Language discrepancy	0.002	0.028**	-0.026*	-0.046**	-0.012	0.032**	-0.015**	-0.003	-0.002	0.011
	(0.008)	(0.012)	(0.015)	(-0.02)	(-0.009)	(-0.016)	(0.006)	(0.011)	(0.017)	(0.017)
Constant	0.070	0.104**	0.037	0.119**	0.03	0.031	0.057	0.096**	-0.095	-0.148*
	(0.056)	(0.045)	(-0.057)	(-0.052)	(-0.049)	(-0.05)	(0.048)	(0.045)	(0.094)	(0.087)
Observations	44,135	34,671	38,490	27,612	40,036	30,712	48,084	38,627	18,865	18,644
R-squared	0.011	0.021	0.017	0.022	0.013	0.022	0.012	0.022	0.028	0.025

Note: Each column contains coefficients of the independent variable in that row on the main dependent variable, grade repetition (which is a binary variable, taking value of 1 if the child was repeating the same grade as last year). All models are estimated using the linear probability model, and include additional controls for religion and social group, whose coefficients are not reported here. Sample weights are applied and state, medium of instruction, and language spoken at home fixed effects are included. Standard errors clustered at the level of the household reported in parentheses; *** p<0.01, ** p<0.05, * p<0.1.

Next, in line with the literature on English training skills (e.g., Azam et al., 2013; Jain et al., 2019), we examine whether the positive or negative associations between language discrepancy is dominant in households where the child goes to an English-medium school. Using the language of instruction and language fixed effects, the model takes English as the base category. In Table 4b (columns 1-2), we see potential evidence of a subgroup in our sample where language capital matters. For females that have a different language spoken at home from their (non-English) medium school, there is 2.8 percentage point increase in the likelihood of grade repetition, a result that is statistically significant at the 5% level. There is no similar association found for boys, implying that language discrepancy could be especially detrimental for girl students who do not go to English-medium schools and speak a different language at home. This is precisely the type of diversity that might be threatened with the imposition of mandatory learning of the home state's language. Similarly, columns 3-6 in Table 4b show the result when we split the sample into Hindi and non-Hindi speaking states.[6] The results further validate the idea that language discrepancy is associated with grade repetition among girls in non-Hindi speaking states. In fact, when we look at the results for Hindi-speaking states, we find a negative association between language discrepancy and grade repetition, with a 2.6 and 4.6 percentage point reduction in the likelihood of grade repetition for boys and girls, respectively. This could also be motivated by existing work that suggests benefits of additive bilingualism (Posel and Casale, 2011) and how learning in one language at home transfers to other (similar) languages as well. These findings all point toward the fact that any adverse effects of language discrepancy are likely to be amplified in non-Hindi speaking states, in non-English medium schools, and largely for girls.

[6] The non-Hindi speaking states are drawn from an official government response (Government of India, 2005). These are Andhra Pradesh, Arunachal Pradesh, Assam, Goa, Jammu & Kashmir, Karnataka, Kerala, Maharashtra, Manipur, Meghalaya, Mizoram, Nagaland, Orissa, Punjab, Sikkim, Tamil Nadu, Tripura, and West Bengal.

Finally, we also look at whether these results are coming from school-going children or university-going students (Columns 7-10, Table 4b). One way to do this is to use the course currently attending data from NSS, which does not distinguish between primary, secondary, and high school, but does differentiate between school and higher secondary (i.e., up to 10th grade, and 12th grade). Our results show that the negative association between language discrepancy and grade repetition among boys is only observed in the school-going sample (up to 10th grade), and disappears when we look at higher secondary school-goers. In line with the main results, there are no statistically significant results for girls.

These findings, when taken together, provide some insight into the sample that might be driving our main result. Boys residing in wealthier households, primarily in urban areas are the main subsample within which language discrepancy is associated with a lower likelihood of grade repetition. In contrast, girls residing in non-Hindi speaking states (e.g., the western state of Maharashtra or the Southern state of Andhra Pradesh) attending schools where the language of instruction is not English are most likely to be facing the adverse effects of language discrepancy in terms of grade repetition. This is in line with similar district-level evidence found in Jain (2017), where he shows that in South Indian states, the mismatch between mother tongue and medium of instruction in schools in so-called “minority” districts is associated with a lower literacy, and educational completion rates than in the districts where there is more congruence between mother tongue and medium of instruction in schools. Our result extends this finding to other non-Hindi speaking states, and is able to make the claim using child-level data, pinpointing that there is a gender-variegated effect.

We also ran a series of robustness checks where we used a logistic regression framework instead of the LPM, and found qualitatively similar results to our main specification. We do not repeat the subgroup analysis using logit, as we consider these results robust to the specification. To check for the sensitivity of the results to the construction of the language discrepancy variable, we also consider an analysis using linguistic distance between the language spoken at home and the medium of instruction in line with Jain (2017).[7] This helps more clearly delineate the *intensity* of the language discrepancy, and any potential (dis)advantages from concordance between the language spoken at home and the medium of instruction at school.

Results from the linguistic distance variable are similar to the overall results, in that there is a statistically significant and negative association between linguistic distance and grade repetition, but only for boys (Figure 1). This suggests that our main results are indeed not sensitive to the definition of the language discrepancy variable, although the sample for this estimation is much smaller, owing to lack of data on linguistic distance to English as well as other languages documented in the NSS. There is no other statistically significant result, further highlighting that language discrepancy's positive impacts on educational outcomes might be driven by the English-medium school-going sample for boys, whereas its negative impacts are more pronounced in the non-English-medium schools for girls. We also used the linguistic distance variable to test the heterogeneity across school-going age children vs. higher-secondary students, finding null results (i.e., there was no statistically significant association between linguistic capital and grade repetition) in this sample.

[7] Table 6 in Jain (2017) provides the linguistic distance matrix for 16 languages in India. The NSS dataset on languages however differs from this in two respects: one, Jain (2017) provides linguistic distance for Rajasthani, which is not recorded as a language in the NSS data. Two, there is linguistic distance between Bihari and other languages in Jain (2017), but the NSS dataset does not include Bihari, but instead has Maithili, a dialect spoken widely in north Bihar. We use this to match the datasets. Notably, data on the distance between these Indian languages and English is absent, so our analysis is restricted to the non-English medium of instruction schools, which are more prevalent in rural India.

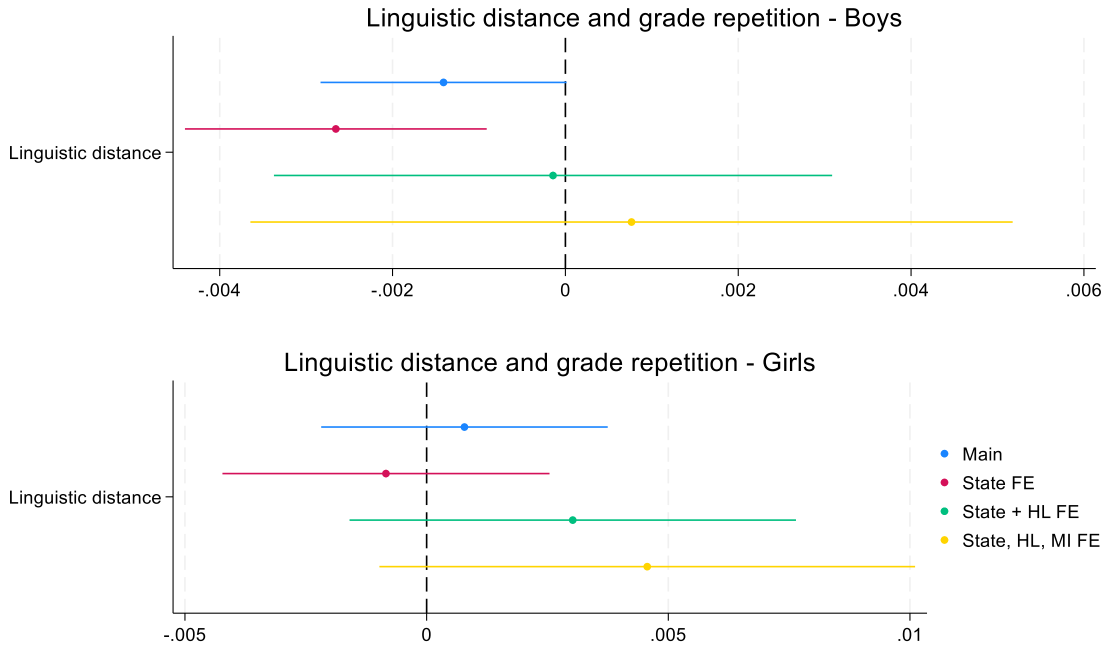

Figure 1: Coefficient plot of linguistic distance on grade repetition for boys and girls

Note: Plot depicts point estimates of coefficients and 95% confidence intervals from four linear probability model regressions of grade repetition on linguistic distance between language spoken at home and medium of instruction at school. HL FE: language spoken at home fixed effects, MI FE: Medium of instruction at school fixed effects. All regression specifications contain additional covariates as reported in Table 3.

6. Conclusions

This work set out to examine the impact of the discrepancy between language spoken at home and the language of instruction at school on grade repetition in the Indian context. In our most robust estimates, we find a statistically significant and *negative* association between language discrepancy and grade repetition for boys, a result that is driven largely by urban households, in Hindi-speaking states, and in relatively wealthier households. In contrast, we find a statistically significant and negative association between language discrepancy and grade repetition for girls, but only for those who study in non-English medium schools or reside in non-Hindi speaking states. These results are robust to a change in the estimation method as well as the definition of the key independent variable.

It is therefore possible that there is a very specific subgroup of children that are most likely to be affected by policies such as the NEP that could potentially impose mother tongue instruction. Such change in policies have been studied extensively (see Ginsburgh & Weber, 2020), but the results are often mixed, pointing to the importance of cultural context in the impact of language of education policies. Our work provides an important exploration of the implication of policies such as NEP, which has already begun implementation in a few states in India as of 2023. The challenge lies in how they will take into account migrant groups in both urban and rural areas, given that they may not have the home language advantage that natives might enjoy when it comes to teaching in the native language.

To be sure, there are several caveats that one should note about the current study. First, it lacks a representative measure of child educational achievement, that is typically measured using cognitive or test scores. The absence of such data at the child level means that we relied on an alternate measure of educational achievement, grade repetition. Past work in grade repetition (e.g., Siddhu, 2011) is primarily focused on the determinants of repetition related to parental involvement or dropping out of the school system (Paul et al., 2021), which we have not examined here. Future work can use other large-scale household survey data to more closely look at the states where language discrepancy could affect educational outcomes. Second, we do not have detailed school-level data on the institutional characteristics. This would be helpful to control for other factors at the school-level that could also be influencing grade repetition, such as grading policies. Third, in developing a robustness check using linguistic distance, we are able to verify that the results are not sensitive to the formulation of the language discrepancy variable. However, this analysis does not account for the linguistic distance between English and Indian regional languages, which would expand the sample considerably. Once such data becomes available, it becomes important to add it to the sample and iterate the analyses to better understand the *intensities* that matter for educational outcomes. Lastly, our dataset is cross-sectional and therefore not able to track any changes that take place over time. Research suggests that linguistic skills can take time to form (Derwing & Munro, 2013), and it is possible that the effects of language discrepancies may change or alter over time for the same children.

Acknowledgments

The authors are grateful to Hansika Kapoor for comments and suggestions, and the editors of the special issue, Tariq Elyas and Donia Smaali Bouhlila. All errors are solely attributable to the authors.

References

Angrist, J., Chin, A., & Godoy, R. (2008). Is Spanish-only schooling responsible for the Puerto Rican language gap? *Journal of Development Economics*, *85*(1–2), 105–128.

Angrist, J. D., & Lavy, V. (1997). The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco. *Journal of Labor Economics*, *15*(1, Part 2), S48–S76.

Argaw, B. A. (2016). Quasi-Experimental Evidence on the Effects of Mother Tongue-Based Education on Reading Skills and Early Labour Market Outcomes. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2760337

Azam, M., Chin, A., & Prakash, N. (2013). The returns to English-language skills in India. *Economic Development and Cultural Change*, *61*(2), 335-367. Retrieved from https://shareok.org/bitstream/handle/11244/321264/oksd_azam_thereturnstoeng_2013.pdf?sequence=1.

Derwing, T. M., & Munro, M. J. (2013). The Development of L2 Oral Language Skills in Two L1

Groups: A 7-Year Study. *Language Learning*, *63*(2), 163–185. https://doi.org/10.1111/lang.12000

Eriksson, K. (2014). Does the language of instruction in primary school affect later labour market outcomes? Evidence from South Africa. *Economic History of Developing Regions*, *29*(2), 311–335. https://doi.org/10.1080/20780389.2014.955272

Ginsburgh, V., & Weber, S. (2020). The Economics of Language. *Journal of Economic Literature*, *58*(2), 348–404. https://doi.org/10.1257/jel.20191316

Government of India. (2018). *National Sample Survey Office Socioeconomic Survey Schedule 25.2: Household Social Consumption: Education* [Data set].

Government of India. (2020). *Unified District Information System for Education Plus (UDISE+) 2019-20*. Department of School Education and Literacy, Ministry of Education. https://dashboard.udiseplus.gov.in/assets/images/pdf/UDISE+2019_20_Booklet.pdf

Ikeda, M., & García, E. (2014). Grade repetition: A comparative study of academic and non-academic consequences. *OECD Journal: Economic Studies*, *2013*(1), 269-315. https://doi.org/10.1787/19952856

Jain, T. (2017). Common tongue: The impact of language on educational outcomes. *The Journal of Economic History*, *77*(2), 473-510. https://doi.org/10.1017/S0022050717000481

Jha, S. (2022, August 22). NEP 2020: mother tongue or no, schools remain tossed between the language of education. *The Financial Express*. Retrieved July 5, 2023, from https://www.financialexpress.com/education-2/nep-2020-mother-tongue-or-no-schools-remain-tossed-between-the-language-of-education/2638825/.

Kabay, S. (2016). Grade repetition and primary school dropout in Uganda. *Harvard Educational Review*, *86*(4), 580-606. https://doi.org/10.17763/1943-5045-86.4.580

Kapur, S., & Chakraborty, T. (2008). English language premium: Evidence from a policy experiment in India. Washington University in St. Louis unpublished paper. Retrieved from: http://www.isid.ac.in/~pu/conference/dec_08_conf/Papers/ShilpiKapur.pdf

Khanuja, S., Bansal, D., Mehtani, S., Khosla, S., Dey, A., Gopalan, B., ... & Talukdar, P. (2021). MuRIL: Multilingual Representations for Indian Languages. *arXiv e-prints*, arXiv-2103. https://doi.org/10.48550/arXiv.2103.10730

Ministry of Human Resource Development, Government of India. (2020). *National Education Policy 2020*. https://www.education.gov.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf

Ministry of Human Resource Development, Government of India. (1968). *National Policy on Education, 1968*. https://www.education.gov.in/sites/upload_files/mhrd/files/document-reports/NPE-1968.pdf

Ministry of Human Resource Development, Government of India. (1992). *National Policy on Education 1986 - Programme on Action 1992*. https://www.education.gov.in/sites/upload_files/mhrd/files/document-reports/NPE-1968.pdf

Nag, S., Vagh, S. B., Dulay, K. M., & Snowling, M. J. (2018). Home language, school language and children's literacy attainments: A systematic review of evidence from low- and middle-income countries. *Review of Education*, rev3.3130. https://doi.org/10.1002/rev3.3130

Paul, R., Rashmi, R., & Srivastava, S. (2021). Does lack of parental involvement affect school dropout among Indian adolescents? Evidence from a panel study. *PLOS ONE*, *16*(5), e0251520. https://doi.org/10.1371/journal.pone.0251520

Saraswathy, M. (2021, July 29). National education policy: one year of steady reforms, a few more miles to go. *Moneycontrol*. Retrieved July 4, 2023, from https://www.moneycontrol.com/news/business/economy/national-education-policy-nep-2020-one-year-steady-reforms-miles-to-go-7242741.html.

Siddhu, G. (2011). Who makes it to secondary school? Determinants of transition to secondary schools in rural India. *International Journal of Educational Development*, *31*(4), 394–401. https://doi.org/10.1016/j.ijedudev.2011.01.008

Seid, Y. (2017). The impact of learning in mother tongue first: Evidence from a natural experiment In Ethiopia. *International Growth Center, LSE, Addis Ababa, Ethiopia*, *3*(2), 400-450. Retrieved from https://riseprogramme.org/sites/default/files/inline-files/Seid_The%20Impact%20of%20Learning%20in%20Mother%20Tongue%20First.pdf

Thrasher, A. W. (1996). Language, Ethnicity, and Regionalism. In J. Heitzman & R. L. Walden (Eds.), *India: A Country Study* (Fifth, pp. 179–197)., Federal Research Division, Library of Congress. Retrieved from the Library of Congress, https://www.loc.gov/item/96019266/

Tsimpli, I. M., Vogelzang, M., Balasubramanian, A., Marinis, T., Alladi, S., Reddy, A., & Panda, M. (2020). Linguistic diversity, multilingualism, and cognitive skills: A study of disadvantaged children in India. *Languages*, *5*(1), 10. https://doi.org/10.3390/languages5010010

Appendix

Table A.1: Summary of languages spoken at home and medium of instruction

Language	Proportion speaking at home	SD speaking at home	Proportion attending school with instruction in	SD attending school with instruction in
Hindi	0.46	0.50	0.44	0.50
English	0.00	0.04	0.27	0.44
Assamese	0.01	0.12	0.02	0.13
Bengali	0.08	0.27	0.07	0.25
Bodo	0.00	0.04	0.00	0.02
Dogri	0.00	0.05	0.00	0.01
Gujarati	0.04	0.19	0.03	0.18
Kannada	0.03	0.18	0.02	0.15
Kashmiri	0.00	0.06	0.00	0.02
Konkani	0.00	0.04	0.00	0.01
Maithili	0.02	0.12	0.00	0.01
Malayalam	0.03	0.16	0.01	0.09
Manipuri	0.00	0.05	0.00	0.01
Marathi	0.07	0.25	0.05	0.22
Nepali	0.00	0.04	0.00	0.02
Oriya	0.03	0.17	0.03	0.16
Punjabi	0.02	0.15	0.01	0.09
Sanskrit	0.00	0.01	0.00	0.01
Santhali	0.00	0.06	0.00	0.00
Sindhi	0.00	0.03	0.00	0.02
Tamil	0.05	0.22	0.03	0.16
Telegu	0.06	0.24	0.02	0.15
Urdu	0.01	0.12	0.01	0.08

Others	0.07	0.25	0.00	0.03
Number of observations			152823	