Effect of State's Language Policy on Learning Outcomes of Linguistic Minorities: Evidence from Indian Districts

Ashokankur Datta* and Anushree Khatri†

September 15, 2023

Abstract

This paper examines how the language policies of linguistic states in India influence the educational outcomes of linguistic minorities. Using the Linguistic Census of India (2011) and the nationally representative ASER dataset, we explore how linguistic heterogeneity within the Indian districts affects the learning outcomes of school-age children. We find that greater linguistic diversity within districts has an adverse impact on the reading scores, with no significant impact on the math scores. These effects are more pronounced among children of parents with lower levels of education, possibly due to limited support. The eastern region, known for its linguistic diversity, experiences more accentuated effects. To conduct our analysis, we employ instrumental variable estimation using a specific attribute of the district's boundary as an instrument. This paper addresses the ongoing debate on the need for decentralizing decision-making to cater to the educational needs of a linguistically diverse nation.

Keywords: Linguistic Minority; Learning Outcome; Language
JEL Codes: I21; J15; O15

*Shiv Nadar Institution of Eminence, Uttar Pradesh, INDIA-201314. E-mail: ashokankur.datta@snu.edu.in
†Shiv Nadar Institution of Eminence, Uttar Pradesh, INDIA-201314. E-mail: ak510@snu.edu.in

1 Introduction

'It is axiomatic that the best medium for teaching a child is his mother tongue. Psychologically, It is the system of meaningful signs that in his mind works automatically for expression and understanding. Sociologically, it is a means of identification among the members of the community to which he belongs. Educationally, he learns more quickly through it than through an unfamiliar linguistic medium'

The Use of Vernacular Languages in Education: The Report of the UNESCO Meeting of Specialists, 1951 as cited in Fishman (1968)

The importance of primary education in one's own mother tongue has long been acknowledged by linguists, educationists, and policymakers as evidenced by the observation of the 70-year-old UNESCO report cited above. Yet, approximately 40% of the world's population does not have access to education in a language they speak or understand (UNESCO, 2016). Such a mismatch is partially a legacy of colonial rule in Asia, Africa and South America, where colonial languages of dominance continue to be the language of administration and power even after political decolonization. Commenting on the educational infrastructure of sub-Saharan Africa, Komarek (1996) writes '... the power elite have maintained the colonial languages, of which only they are masters, as a way of ensuring their access to information, and thus preventing self-determination and sharing of power by others.' In addition, the birth of linguistic nation-states and sub-national provinces has often led to the marginalization of minority and indigenous languages by dominant linguistic groups. Linguistic homogenization has often been thought of as a way towards national unity and integration (Dutcher, 2001). Thirdly, there are cost concerns regarding the provisioning of education in multiple mother tongues in linguistically diverse societies of Asia and Africa. In addition to such supply-side constraints, parents often make choices without considering the difficulties children face in learning a second language, in addition to their mother tongue, for academic purposes. Such decisions are often based on an understanding that does not distinguish between the process of learning a second language and the process of learning through a second language (Dutcher, 2001). Numerous studies have estimated how such a mismatch between the first language/ mother tongue and the medium of instruction affects the learning outcomes of children (Thomas and Collier, 1997; Angrist et al., 2008; Taylor and von Fintel, 2016; Jain, 2017;

Bernhofer and Tonin, 2022).

India, like the rest of South Asia, is a land of great linguistic diversity. There is a saying in Hindustani that suggests the spoken language in this region changes every 7 kilometers [1]. While the colonial government in India started a process of vernacularization of administration in 1837, it created a hierarchy among Indian languages based on the 'viability of a language being an official language' (Mishra, 2020). The accreditation of fourteen Indian languages as scheduled languages by the Indian Constitution in 1950 and the linguistic reorganization of states six years later, formalized the hierarchical distinction between Indian languages and their access to state patronage[2]. After the linguistic reorganization of states in 1956, most states adopted the language spoken by the majority population as the official language. Many of these states had a large share of the population that didn't speak the official language as their first language[3]. Any empirical analysis, that focuses exclusively on the scheduled languages and an aggregated heterogeneous category called 'others', ignores the immense linguistic heterogeneity and substantial presence of linguistic minorities within each of India's linguistic states. The Census of India 2011 reports the presence of 19569 mother tongues which through a process of 'rationalization', is aggregated to 22 scheduled languages and (99+1) non-scheduled languages. The number of languages that act as the official language of any Indian state is less than even the number of scheduled languages. Thus, a large number of Indians speak non-scheduled languages and/or do not speak the official language of their state as their first language.

The paper tries to understand the effect of linguistic minority status on the learning outcomes of Indian children. More precisely we ask the question: how does the higher

[1]The saying in Hindustani is *'Kos kos par pani badley, char kos par vani'* which roughly translates to *The taste of water changes after every 1.8 km, the spoken word changes after 7.2 kms'*

[2]In 1950, the Constitution of India recognized 14 Indian languages as scheduled languages in the 8th Schedule. Eight other Indian languages were added to this list through constitutional amendments till 2004. Being a scheduled language ensures that the state is obligated to take measures for the development of the language.

[3]According to the 2011 census, approximately 66% of people in Karnataka do not report Kannada as their mother tongue. This proportion has remained unchanged since 1971. In Assam, approximately 48% of people reported Assamese as their mother tongue in 2011. This has declined from 56% in 1951. However, it should be noted Bengali is used as an official language in three districts of Assam.

proportion of linguistic minorities in a district affect the learning outcomes of children? While in an ideal situation, one would like to study the impact of an individual's linguistic minority status on learning, the lack of detailed linguistic information in surveys with educational information requires us to modify the question and define the explanatory variable of interest at a district level. This paper contributes to a large literature that studies the impact of linguistic minority status on learning outcomes. Since the medium of instruction in public schools in India is usually the scheduled official language, it also speaks to the literature that studies the impact of not having one's first language as the medium of instruction in schools. Most of the papers in this literature are situated in Sub-Saharan Africa and South American contexts. Using longitudinal data from South Africa, Taylor and von Fintel (2016) show that mother tongue instruction in lower grades improves the academic performance of students in later grades. Using a natural experiment in Morocco that changed the medium of instruction in Morocco from French to Arabic in 1983, Lavy (1997) show that the introduction of Arabic medium of instruction reduced the French language skills in later years. However Angrist et al. (2008) do not find similar results in their study of Puerto Rican schools. Using the change in language policy under the Bantu Education Act of 1953, Eriksson (2014) find that greater exposure to mother tongue instruction in primary schools positively affects literacy, years of education, and numerical literacy. In addition to such causal studies, there exist numerous qualitative studies by educationists and linguists that suggest a positive impact of mother tongue instruction on learning outcomes (Hakuta et al., 2000; Brock-Utne, 2007; Piper and Miksic, 2001).

The literature on the effect of linguistic minority status or medium of instruction on learning outcomes in India is relatively scarce. Jain (2017) is a notable exception. Using the 1956 linguistic reorganization of states in South India, Jain (2017) estimates the effect of a mismatch between mother tongue and official language on outcomes like literacy, school completion, and matriculation rates. While similar to Jain (2017) in terms of the question of interest, this paper differs in a number of significant ways. Firstly, Jain (2017) restricts its geographical area of interest to South India since this region experienced a major exchange of areas between states during the 1956 linguistic reorganization. Our study considers all regions of India. This is especially important since North East India

and the Eastern Plateau region constitute the most linguistically diverse regions in the country. Secondly, Jain (2017) consider coarse measures of learning (literacy, school enrollment, and matriculation rates) as dependent variables. We consider actual test scores from a language and mathematics test conducted all over India by the educational NGO PRATHAM as a part of a nationally representative survey. It should also be mentioned that ours is one of the few studies that also studies the impact of language status on math learning. In most papers mentioned above, with the exception of Lavy (1997), literacy and language skills are the outcomes of interest. Thirdly, Jain (2017) uses the list of 22 scheduled languages to define linguistic minorities and diversity. This grossly understates the linguistic diversity of a region. We use a list of 122 languages to define the concept of linguistic minorities.

In our paper, we employ fixed effect panel regressions and a two-stage least squares (2SLS) analysis to identify the impact of linguistic minority status on learning outcomes. Our IV results indicate a positive effect of linguistic minority status on read scores but not on math scores. Based on our preferred specification in the IV framework, we find that one standard deviation increase in the Herfindahl Index and the percentage of official language speakers of a district increases the read score by 0.003 standard deviations and 0.0035 standard deviations, respectively. Children whose parents lack primary education experience a decline in reading scores when located in linguistically heterogeneous districts. Furthermore, our analysis highlights notable regional disparities, with the eastern region being characterized as the most linguistically diverse part of the country, exhibiting a more pronounced effect than other regions.

The rest of the paper is structured as follows: Section 2 describes the data used, Section 3 describes the empirical strategy, Section 4 reports the results, and Section 5 concludes.

2 Data

The empirical analysis in this paper uses data from multiple sources. First, we obtain rich information on the educational outcomes of children from six rounds (2011-14,

2016, and 2018)[4] of the Annual Status of Education Report (ASER) dataset, a nationally representative cross-section household survey. Each round of the ASER dataset provides information on children's schooling status, foundational reading and numeracy knowledge level, and additional information on parents and household characteristics. The survey employs a two-stage sampling strategy: it selects 20 villages using Probability Proportional to Size (PPS), then employs simple random sampling to choose 30 households per rural district in India. With coverage of 600 households per district, the survey aggregates data from around 300,000 households annually, encompassing roughly 600,000 children aged 3 to 16.

ASER is a distinct survey, conducted at home encompassing out-of-school children and those attending traditional schools like Madrasas. ASER's large sample size and sampling strategy provide representative estimates at the district level. By employing more nuanced and granular learning outcome variables, the ASER dataset is an improvement over other datasets like NSS and Census which predominantly rely on coarser outcomes, primarily focusing on the highest level of education attended. ASER administers its own reading and arithmetic assessments to test all the children, except children of the age 3-4 years, in the regional language[5] of each child to determine the reading and mathematic learning outcomes for the children in the sampled households. The same test is given to children across the years[6]. This distinction allows us to gain a richer and more comprehensive understanding of the impact of linguistic status on learning outcomes.

Ideally, we would have accounted for the linguistic mismatch between a child's mother

[4]ASER survey was not conducted in 2015. In 2017, ASER conducted an alternative survey, ASER Beyond Basics, focusing on children aged 14-18 years in 28 districts across India.

[5]Notably, between 2011-14, the test was conducted in 15 regional languages, with each child tested in their respective native language, and in 2016 and 2018, the test was conducted in 19 regional languages.

[6]During the assessment, ASER surveyors present four questions to each child for both reading and arithmetic tests. The reading test comprises four levels: letters, words, a short paragraph (equivalent to grade 1 text), and a short story (equivalent to grade 2 text). Similarly, the arithmetic assessment consists of four levels: single-digit number recognition, double-digit number recognition, two-digit subtraction with carry over, and three-digit by one-digit division (corresponding to grade 3 or 4 standards). Surveyors begin with the most difficult question and proceed to simpler ones if needed. The highest attained proficiency level in each test is then recorded for each child.

tongue and the medium of instruction in school. This would involve analyzing the impact of such a mismatch on reading and math scores. Unfortunately, ASER does not collect information about the mother tongue of the chosen children. To address this issue, we use the Census of India (2011) to construct linguistic information variables at the district level. The Census of India is specifically designed to collect demographic and socio-economic characteristics at the administrative level. ASER also lacks information on the caste and religion of the children, which significantly influence educational outcomes in India. Therefore, we utilize the Census demographic dataset to compute district-level percentages of the rural population, Muslim population, Scheduled Castes (SCs), and Scheduled Tribes (STs) residing in rural areas. The lowest level of geography aggregated to which the ASER sample can be mapped is the district.

The Linguistic Census of India (2011), part of the larger Census, is the richest source of language data collected and published at successive decennial censuses. It captures detailed language data that encompasses the ethnic and linguistic characteristics of the population. Within the household schedule of the Census questionnaire, there is a dedicated inquiry regarding the mother tongue of each individual residing in the household. In total, the Census yielded 19,569 raw returns of mother tongues. The process of categorizing the 19569 mother tongues into 22 scheduled languages understates the linguistic diversity in India. Researchers using datasets like the National Sample Survey (NSS) and the National Family Health Survey (NFHS) focus more on the 22 scheduled languages. In an effort to comprehensively account for the linguistic diversity in India, our analysis incorporates 122 scheduled and non-scheduled languages, in contrast to the commonly used 22 scheduled languages employed by others. This broader selection of languages allows us to address the concern of linguistic diversity in India more effectively.

Third, we use the district-level administrative boundary of the Census 2011 data shape file. Using this data, we compute the latitude and longitude of the centroid of the district to control for the geographical location of the districts. We also construct a variable that indicates whether a district is a border district of that state[7].Concurrently, the shape

[7]First, we identify the neighbors of the district and then if the district shares its boundary with other districts of some other state, then we say it is a border district.

file and official state languages serve as a foundation for our instrumental variable. To construct this instrumental variable, we first compute the cumulative border length of each district. Next, we determine the neighboring districts for each district and analyze their official languages. This comparative analysis guides the creation of a variable that represents the percentage of a district's total boundary shared with neighboring districts that have a different official language[8]. Table 1 presents the summary statistics for our dependent and independent variables.

3 Empirical Approach

In this section, we lay out an econometric model to identify the causal effect of linguistic minority status on children's learning outcomes. As discussed in the earlier section, data on linguistic status can be merged with the learning outcomes only at the district level. Therefore, our main independent variable is at the district level. By utilizing the Linguistic Census of India, we create two district-level independent variables, drawing from 121 scheduled and non-scheduled languages for enhanced precision in capturing linguistic diversity, compared to the scheduled 22 languages.

The first variable, the Percentage of Official Language Speakers (POLS), signifies the percentage of the total population speaking the official language in a district. We hypothesize that districts with a higher population speaking the official language will have fewer people with a linguistic mismatch. Thus, a lower number of people whose scores would be affected due to the linguistic mismatch. The second variable is measured using the Herfindahl index, which measures the language concentration within the district. A higher Herfindahl index indicates the dominance of certain languages, while a lower value suggests greater linguistic diversity. This index enables us to explore how language concentration or diversity might influence test scores, with lower Herfindahl values potentially

[8]To illustrate, consider Darjiling district in West Bengal, contiguous with six districts - three from Sikkim, two from West Bengal, and one from Bihar. Darjiling's total border length is 267.95 km sq. Given that Bihar's official language is Hindi while Nepali is one of the official languages of both Sikkim and West Bengal, the shared boundary with the language-mismatched district amounts to 46.32 sq km. This translates to Darjiling sharing 17.29% of its border with a district having an official language dissimilar to its own. This percentage constitutes our instrumental variable, labeled as DifferentBoundary in our study.

associated with lower scores due to linguistic heterogeneity and its potential impact on learning outcomes.

3.1 Baseline Model

We begin by estimating the following equation using OLS:

$$Y_{idsr} = \alpha_0 + \alpha_1 Language_{ds} + \alpha_2 X_{idsr} + \alpha_3 Z_{ds} + \alpha_4 W_{ds} + \zeta_{sr} + \varepsilon_{idsr} \quad (1)$$

where, Y_{idsr} are the learning outcomes measured by the read and math scores of child i in district d state s and round r. $Language_{ds}$ serves as the metric for assessing linguistic homogeneity in the district, measured via POLS and the Herfindahl Index. X_{idsr} is a vector of individual level controls such as child sex, child age, household electricity connection, toilet availability, the household has a mobile, household type, the father attended school, and the mother attended school. ζ_{sr} are the state-round fixed effects, controlling for all time-varying state-level factors. We could not control for the district-fixed effects as our main variable $Language_{ds}$ is at the district level. Instead, we rely on a diverse set of district-level control variables. Z_{ds} and W_{ds} are vectors representing the geographical and demographic characteristics of districts, respectively.[9] Standard errors are clustered at the district level.

The main challenge of identifying β from Equation (1) is the possibility that some omitted variables which vary across districts and over time, may be correlated with both $Language_{ds}$ and Y_{idsr}. The presence of time-varying unobservable effects at the district level cannot be ruled out. For instance, districts undergoing significant demographic shifts can impact a district's Herfindahl Index and POLS, subsequently affecting children's learning outcomes through language barriers. Additionally, endogeneity concerns arise in POLS, as the assignment of official language isn't random but is influenced by the predominant language spoken by the state's majority population.

[9]Demographic characteristics of a district include Demographic controls include the percentage of the rural SC population, percentage of the rural ST population, percentage of the rural population, and percentage of the Muslim population. Geographic Controls include latitude, longitude, and boundary district dummy.

3.2 IV Strategy

To address the concern of endogeneity, We use $DifferentBoundary_{ds}$ as an instrument for $Language_{ds}$. It shows the percentage of a district's total boundary shared with neighboring districts having a different official language. Identification in this method relies on the quasi-randomness of

We estimate a two-stage least squares (2SLS) method, which is specified as follows:

$$Language_{ds} = \beta_0 + \beta_1 DifferentBoundary + \beta_2 X_{idsr} + \beta_3 Z_{ds} + \beta_4 W_{ds} + \lambda_{sr} + \eta_{idsr} \quad (2)$$

$$Y_{idsr} = \alpha_0 + \alpha_1 Language_{ds} + \alpha_2 X_{idsr} + \alpha_3 Z_{ds} + \alpha_4 W_{ds} + \zeta_{sr} + \varepsilon_{idsr} \quad (3)$$

Equations (2) and (3) are the second stage and first stage, respectively. Our main explanatory variables, the linguistic minority status, measured through POLS and Herfindahl Index $Language_{ds}$ are instrumented by $DifferentBoundary_{ds}$. Y_{idsr} are the read and math scores of child i in district d in state s and round r. As above, X_{idsr}, Z_{ds}, and W_{ds} are individual, geographic, and demographic level controls, respectively. We use an IV-2SLS procedure and cluster standard errors at the district level.

3.2.1 Validity of instrument

We test whether the instrument,$DifferentBoundary_{ds}$, is a good predictor of the endogenous variable, *Herfindahl Index* or *POLS*, in the first stage of the 2SLS estimation. Table 2 presents the results of the first-stage regressions to test for the relevance condition. We present the results for the relevance condition in the next section.

Boundary districts tend to differ from non-boundary districts in several aspects, potentially making them less developed. These differences could be attributed to factors such as distance from the capital city, limited access to resources and infrastructure, people tend to invest less, and perhaps even historical or political factors. Therefore, we control for boundary in our specification conditional on controlling for boundary dummy. Boundary districts can also affect the learning outcomes directly. However, we control for boundary districts in the structural equation. controlling for boundary districts, we believe that $DifferentBoundary_{ds}$ affects learning outcomes only through POLS and Herfindahl Index. The justification for using $DifferentBoundary_{ds}$ as an instrument lies

in the notion that boundary districts tend to have a more diverse linguistic population than non-boundary districts. This diversity is posited to be a consequence of the geographical location of boundary districts, where people living near district borders have increased linguistic interactions, cultural exchanges, and even migration across linguistic boundaries. Economic prospects, trade, and political considerations can drive migration, leading migrants to bring their languages, and bolstering linguistic diversity. Herfindahl Index is expected to be higher in boundary districts due to the influence of migration and cultural exchange with neighboring regions. Some boundary districts might have historical ties to multiple linguistic and cultural groups. These ties can lead to the preservation of languages even when they are no longer widely spoken elsewhere. In contrast, non-boundary districts tend to harbor more homogenous cultures. The instrument captures the impact of linguistic diversity, influenced by the unique characteristics of boundary districts, which can help control for potential biases and provide a true causal effect of linguistic minority status on learning outcomes.

4 Results

4.1 OLS Results

Table 2 presents the OLS estimates of the impact of linguistic minority status on read and math test scores. Panel A shows the estimates for the read score, and Panel B shows for the math score. Column (1) reports the effect of the *Herfindahl Index* and *POLS* on scores from the regression equations with state-round fixed effects but no controls. Progressively, we add individual, geographic, and demographic controls in columns (2), (3), and (4), respectively.

Our results indicate that belonging to a linguistically homogeneous district increases read scores but does not affect math scores. In our baseline specification in Column (1), we find that, on average, a percentage point increase in the *Herfindahl Index* increases the read score by 0.0041 points and the math score by 0.0037 points. Similarly, a one percentage point increase in POLS results in a 0.0029 point increase in read scores and a 0.0025 percentage point increase in math scores. We find that in naive specifications, the effects are statistically significant. Upon introducing individual and geographic controls,

the effects remain statistically significant. In our preferred specification, including individual controls, geographic controls, demographic controls, and state-round fixed effects, we note that a percentage point increase in the *Herfindahl Index* increases the read score by 0.0009 points. It means that a standard deviation increase in the *Herfindahl Index* increases the read score by 0.0006 standard deviations or 0.02% of the sample mean of the reading score of these children. However, for *POLS*, we find that the effect on read score is not statistically significant. Additionally, we find that the linguistic minority status does not have a statistically significant impact on the math score of the children.

4.2 IV Results

Table 3 presents the IV-2SLS estimates using $DifferentBoundary_{ds}$ as the source of exogenous variation.[10] In panel A, our dependent variable is read score and in panel B our dependent variable is the math score. The first-stage results, estimated using Equation (3), are shown in Columns (1) and (3). It examines whether $DifferentBoundary_{ds}$ predicts the *Herfindahl Index* and *POLS*, which are endogenous regressors, in the absence of control variables and include state-round fixed effects. In Column (1), for read score, we find that $DifferentBoundary_{ds}$ is significantly correlated with the *Herfindahl Index*. We check for the relevance condition by checking for F-statistic. The F-statistic for the regression model is 33.83. Estimating for *POLS*, we find that the F-statistic for the regression model is 36.14. In Column (3), we add individual, geographical, and demographic controls with state-round fixed effects to the structural equation. Despite this augmentation, the coefficients for both *Herfindahl Index* and *POLS* remain negative and significant. Our first-stage results are significant at the 0.1% significance level. The F-statistics for the regression models with these additional controls are 26.89 and 26.61 for *Herfindahl Index* and *POLS*, respectively. We find that a percentage point increase in the border shared with a district that has a different official language decreases the *Herfindahl Index* by 0.23 percentage points and *POLS* by 0.20 percentage points. The first stage results and F-statistics for math score are similar to the read score. The results indicate that a district with a greater percentage of border share with linguistic mismatch districts has a lower Herfindahl Index or POLS.

[10] Standard errors are clustered at the district level in all the specifications.

Moving to Columns (2) and (4), we report the results of the second stage estimated using equation (2). In panel A, the estimated coefficient for the *Herfindahl Index* is positive and statistically significant in both columns, with and without controls. In column (2), we find that a percentage point increase in *Herindahl Index* increases the read score by 0.009 points, while, a one percentage point increase in *POLS* is linked to a 0.011 point increase in the read score. Notably, these estimates are more than twice as large as the OLS estimates reported in Table 2. The effects of *Herfindahl Index* and *POLS* on the read score are highly significant at the 1% level. Additionally, we find that the effect of *Herfindahl Index* and *POLS* on math score is also significant at the 1% level, when considering state-round fixed effects and no controls.

Finally, in Column (4), considering state-round fixed effects and a full set of controls, a one percentage point increase in the *Herfindahl Index* increases the read score by 0.0045 points. Similarly, a one percentage point increase in *POLS* increases the read score by 0.0052 points. It means one standard deviation increase in the *Herfindahl Index* and *POLS* increases the read score by 0.003 standard deviations and 0.0035 standard deviations, respectively. These results indicate that a percent increase in the *Herfindahl Index* *POLS* increases the read score by 0.0012 and 0.0015 percent, respectively. However, the effect of *Herfindahl Index* and *POLS* on math score is smaller in magnitude and is no longer statistically significant after including all the controls. This suggests that being of minority linguistic status does not affect math score to the extent it affects the read score.

We carry out robustness checks to assess the robustness of our results. First, we consider only the major 18 states instead of all states.[11] We find that the results are still positive and statistically significant.[12] Next, to check if our results are driven by any particular ASER round. We one by one run regressions dropping a year in each regression. We find that our results are robust to the exclusion of an ASER round.[13]

[11]Major 18 states are Punjab, Uttarakhand, Haryana, Rajasthan, Uttar Pradesh, Bihar, West Bengal, Jharkhand, Odisha, Chhattisgarh, Madhya Pradesh, Gujarat, Maharashtra, Andhra Pradesh, Kerala, and Tamil Nadu.

[12]We present results for major 18 states in Appendix.

[13]We present coefficient plots for *Herfindahl Index* and *POLS* in the Appendix.

4.3 Heterogeneous Effects

Table 4 presents the heterogeneous impact of linguistic minority status on read score by parental education. The findings reveal an intriguing trend: children whose parents have not completed primary education demonstrate higher read scores when associated with a linguistic minority group, in contrast to those whose parents have completed such education. This suggests a distinct challenge for first-generation learners in areas where their language is not prevalent. In cases where parental formal education is lacking, children often encounter difficulties in comprehending school coursework, struggling to access suitable guidance and support. The language of instruction emerges as pivotal; alignment with the student's language enhances their understanding, cultivating both academic interest and improved performance. Conversely, children from educated families residing in linguistically minority districts benefit from additional resources, aiding their grasp of school subjects.

Next, we try to find the effect of linguistic minority status on read score in Hindi states[14] and four regions of India. Table 5 presents the results for Hindi states and East region states.[15] Hindi states are those states in which one of their official languages is Hindi. We find that the impact of *Herfindahl Index* and *POLS* on the read score is smaller in comparison to when we consider all the states. In the Hindi states, the Herfindahl Index and POLS are higher than in the non-Hindi states. There is less variation in the Herfindahl Index and POLS. Therefore, we see that there is no significant impact on read score if belonging to one of the Hindi states.

Further, we find heterogeneous effects across regions of the linguistic minority status on the read score across the regions and find that East region states[16] exhibit the most pronounced impact on children's scores (Panel B). The higher impact of the Herfindahl Index and POLS on read score in the east region could be because this region has a

[14]The Hindi States are Himachal Pradesh, Uttarakhand, Haryana, Delhi, Rajasthan, Uttar Pradesh, Bihar, Jharkhand, Chhattisgarh, Madhya Pradesh, Gujarat, Dadra & Nagar Haveli, Maharashtra, and Andaman & Nicobar Island

[15]We present results for North, West, and South regions in the Appendix.

[16]East region states are Bihar, Sikkim, Arunachal Pradesh, Nagaland, Manipur, Mizoram, Tripura, Meghalaya, Assam, West Bengal, Jharkhand, Odisha, and Chhattisgarh.

greater variation in the Herfindahl Index and POLS. This variation underscores a diverse linguistic and socio-economic landscape, contributing to a setting where the influence of linguistic minority status on read scores becomes more accentuated. Moreover, multiple official languages within the East region states introduce an additional layer of complexity to the educational environment. The broader array of official languages amplifies the challenge of navigating linguistic diversity, thereby contributing to distinct variations in the impacts on read scores among linguistic minority students. For other regions, we find that our results are not significant and the F-stat is lower than 10. In summation, the study elucidates the distinct impact of linguistic minority status on reading scores across various regions.

5 Conclusion

This study evaluates the effect of linguistic heterogeneity on learning outcomes for school-going children in the Indian context. Employing a fixed effect panel regressions and a two-stage least squares regression model, we show that there is a decline in the reading ability of the children belonging to the linguistically heterogeneous districts. We find that these effects are significant only for the read scores, as no such effect is found on the math scores. This difference highlights the difficulty caused by belonging to the linguistically heterogeneous district, especially for subjects like reading and writing, which are fundamental for academic success. We further show that these effects are higher for children whose parents have not completed primary education, possibly due to limited support and resources at home. Regionally, the Eastern part of India, known for its linguistic diversity, experiences more pronounced effects. Based on our research and existing literature, enhancing educational outcomes in linguistically diverse regions of India can potentially be achieved by implementing native language instruction and decentralizing language instruction decisions.

References

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

Bernhofer, J. and M. Tonin (2022): "The effect of the language of instruction on academic performance," *Labour Economics*, 78.

Brock-Utne, B. (2007): "Language of instruction and student performance: new insights from research in Tanzania and South Africa," *International Review of Education*, 53, 509–530.

Dutcher, N. (2001): "Expanding Educational Opportunity in Linguistically Diverse Societies," Center for Applied Linguistics, Washington, DC.

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.

Fishman, J. A., ed. (1968): *THE USE OF VERNACULAR LANGUAGES IN EDUCATION: THE REPORT OF THE UNESCO MEETING OF SPECIALISTS, 1951*, Berlin, Boston: De Gruyter Mouton, 688–716.

Hakuta, K., Y. Butler, and D. Witt (2000): "How long does it take English learners to attain proficiency?" Technical report policy. Report No. 2000-1 University of California Linguistic Minority Research institute, Berkeley.

Jain, T. (2017): "Common tongue: The impact of language on educational outcomes," *The Journal of Economic History*, 77, 473–510.

Komarek, K. (1996): "Mother-tongue education in sub-Saharan countries: Conceptual and strategic considerations for the promotion of mother-tongue education in Africa." Eschbom, Germany: German Technical Cooperation Agency. Mimeo.

Lavy, J. D. A. . V. (1997): "The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco." *Journal of Labor Economics*, 15(1).

Mishra, P. (2020): *In Language and the Making of Modern India: Nationalism and the Vernacular in Colonial Odisha, 1803–1956*, Cambridge: Cambridge University Press., chap. How the Vernacular Became Regional.

Piper, B. and E. Miksic (2001): "Mother tongue and reading: Using early grade reading assessments to investigate language-of-instruction policy in East Africa," in *The early grade reading assessment: Applications and interventions to improve early grade literacy*, ed. by A. Gove and A. Wetterberg, RTI Press.

Taylor, S. and M. von Fintel (2016): "Estimating the impact of language of instruction in South African primary schools: A fixed effects approach," *Economics of Education Review*, 50, 75–89.

Thomas, W. P. and V. Collier (1997): "School Effectiveness for Language Minority Students," NCBE Resource Collection Series, No. 9. National Clearinghouse for Bilingual Education, Washington, DC.

UNESCO (2016): "If you don't understand, how will you learn?" Global Education Monitoring Report, Policy Paper 24.

Table 1: Summary Statistics

	Observations (1)	Mean (2)	SD (3)
Outcomes			
Read Score	2568003	3.546	1.504
Math Score	2561404	3.279	1.329
Individual Characteristics			
Female	3370894	0.478	0.499
Child Age	3391555	9.661	3.768
Father Attended School	3129895	0.737	0.440
Mother Attended School	3306600	0.532	0.499
Household Characteristics			
Has Electricity Connection	3769725	0.797	0.402
Has Toilet	3758265	0.497	0.499
Has Mobile	3715941	0.753	0.431
Puca House	3763480	0.404	0.491
Semi-Puca House	3763480	0.272	0.445
District Characteristics			
Latitude	640	23.413	5.809
Longitude	640	81.05	6.28
Boundary District	631	0.583	0.493
Percentage of SCs living in Rural	631	15.984	10.362
Percentage of STs living in Rural	631	20.052	28.802
Percentage of Rural Population	631	73.603	21.118
Percentage of Muslim	640	12.862	17.412
Herfindahl Index	640	0.734	0.227
POLS	640	78.747	29.141

Notes: Author's calculation using ASER and Census of India dataset. The unit of observation is district in "District Charachteristics".

Table 2: OLS Estimates of the Effect of Herfindahl Index and POLS on Learning Outcomes

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
				Panel A: Dependent Variable is Read Score				
Herfindahl Index	.4167112***	.1783808***	.150746***	.0965252**				
	(0.0669)	(0.0420)	(0.0416)	(0.0472)				
POLS					.0028918***	.0007561**	.0008082**	.0002451
					(0.0007)	(0.0003)	(0.0003)	(0.0004)
Observations	2565742	2250433	2232047	2232047	2565742	2250433	2232047	2232047
				Panel B: Dependent Variable is Math Score				
Herfindahl Index	.3750981***	.1547842***	.1349821***	.0280863				
	(0.0632)	(0.0411)	(0.0410)	(0.0449)				
POLS					.0024872***	.0004902	.0005338	-.0004904
					(0.0007)	(0.0003)	(0.0004)	(0.0003)
Observations	2559145	2245433	2227077	2227077	2559145	2245433	2227077	2227077
State Round Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Geographic Controls	No	No	Yes	Yes	No	No	Yes	Yes
Demographic Controls	No	No	No	Yes	No	No	No	Yes

Notes: This table presents the coefficients estimated for read score and math score in children aged 5-16 years through two specifications: (1) utilizing the Herfindahl Index as the independent variable (Columns 1-4), and (2) using POLS as the independent variable (Columns 5-8). Panel A shows estimates for read score, and Panel B for math score. We include state-round fixed effects in all the specifications. State refers to the state of residence at the time of the survey. Round refers to the survey round in which the child was interviewed. Additionally, subsequent columns incorporate various controls: Individual controls encompassing child sex, child age, household electricity connection, toilet availability, the household has a mobile, household type, the father attended school, and the mother attended school. Geographic Controls include latitude, longitude, and boundary district dummy. Demographic controls include the percentage of SCs population living in rural, percentage of STs population living in rural, percentage of the rural population, and percentage of the Muslim population. Standard errors are in parentheses and are clustered at the district level in all the specifications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: 2SLS Estimates of the Effect of Herfindahl Index and POLS on Read Score

	(1) First Stage	(2) Second Stage	(3) First Stage	(4) Second Stage	(1) First Stage	(2) Second Stage	(3) First Stage	(4) Second Stage
			Panel A: Dependent Variable is Read Score					
Different Boundary	-.0029117***		-.002338***		-.2340749***		-.2006174***	
	(0.0005)		(0.0005)		(0.0389)		(0.0389)	
Herfindahl Index		.9016817***		.4454187**				
		(0.2361)		(0.2108)				
POLS						.0112164***		.0051909**
						(0.0031)		(0.0026)
Observations	2540183	2540183	2227535	2227535	2540183	2540183	2227535	2227535
F-statistic	33.83075		26.89547		36.1491		26.61107	
			Panel B: Dependent Variable is Math Score					
Different Boundary	-.0029117***		-.0023381***		-.2340464***		-.2005584***	
	(0.0005)		(0.0005)		(0.0389)		(0.0389)	
Herfindahl Index		.6487244***		.1517121				
		(0.2231)		(0.2013)				
POLS						.0080707***		.0017687
						(0.0029)		(0.0024)
Observations	2533658	2533658	2222587	2222587	2533658	2533658	2222587	2222587
F-statistic	33.8781		26.92868		36.14606		26.60057	
State Round Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual Controls	No	No	Yes	Yes	No	No	Yes	Yes
Geographic Controls	No	No	Yes	Yes	No	No	Yes	Yes
Demographic Controls	No	No	Yes	Yes	No	No	Yes	Yes

Notes: This table presents the coefficients estimated for read score and math score in children aged 5-16 years through two specifications: (1) utilizing the Herfindahl Index as the independent variable, and (2) using POLS as the independent variable. Panel A presents estimates for read score, and Panel B for math score. We include state-round fixed effects in all the specifications. State refers to the state of residence at the time of the survey. Round refers to the survey round in which the child was interviewed. Column (1) and (3) presents the first stage results and Column (2) and (4) presents the second stage results. Column (1)-(2) presents results without controls and Column (3)-(4) presents with controls. Individual controls encompass child sex, child age, household electricity connection, toilet availability, the household has a mobile, household type, the father attended school, and the mother attended school. Geographic Controls include latitude, longitude, and boundary district dummy. Demographic controls include the percentage of SCs population living in rural, percentage of STs population living in rural, percentage of the rural population, and percentage of the Muslim population. Standard errors are in parentheses and are clustered at the district level in all the specifications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Heterogeneous Effect of Linguistic Minority Status by Parental Education on Read Score

	(1) First Stage	(2) Read Score	(3) First Stage	(4) Read Score
Panel A: Both Parents Primary Educated				
Different Boundary	-.0023302***		-.1995219***	
	(0.0004)		(0.0345)	
Herfindahl Index		.4259316**		
		(0.1837)		
POLS				.0049744**
				(0.0022)
Observations	918398	918398	918398	918398
F-statistic	28.90178		33.43592	
Panel B: Both Parents Primary Uneducated				
Different Boundary	-.0021863***		-.1825158***	
	(0.0005)		(0.0488)	
Herfindahl Index		.4513378		
		(0.2935)		
POLS				.0054063
				(0.0037)
Observations	593370	593370	593370	593370
F-statistic	16.19687		14.00069	
State Round Fixed Effect	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes
Geographic Controls	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes

Notes: This table depicts heterogeneous effects on the read scores of children aged 5-16 based on parental education. Two specifications are employed: (1) utilizing the Herfindahl Index in Columns 1-2, and (2) using POLS in Columns 3-4. Panel A presents results for Both Parents Primary Educated. Panel B presents results for Both Parents Primary Uneducated. Individual controls encompass child sex, child age, household electricity connection, toilet availability, the household has a mobile, household type, the father attended school, and the mother attended school. Geographic Controls include latitude, longitude, and boundary district dummy. Demographic controls include the percentage of SCs population living in rural, percentage of STs population living in rural, percentage of the rural population, and percentage of the Muslim population. Standard errors are in parentheses and are clustered at the district level in all the specifications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Heterogeneous Effect of Linguistic Minority Status For Hindi States and Regions on Read Score

	(1) First Stage	(2) Read Score	(3) First Stage	(4) Read Score
Panel A: Hindi States				
Different Boundary	-.0029921***		-.4138622***	
	(0.0009)		(0.0995)	
Herfindahl Index		.3171899		
		(0.2478)		
POLS				.0022932
				(0.0017)
Observations	1400061	1400061	1400061	1400061
F-statistic	11.33906		17.30505	
Panel B: East Region				
Different Boundary	-.002919***		-.3095609***	
	(0.0007)		(0.0842)	
Herfindahl Index		.8671797***		
		(0.3052)		
POLS				.0081771***
				(0.0031)
Observations	746608	746608	746608	746608
F-statistic	16.76006		13.51527	
State Round Fixed Effect	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes
Geographic Controls	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes

Notes: This table depicts heterogeneous effects on the read scores of children aged 5-16 for Hindi states and regions. Two specifications are employed: (1) utilizing the Herfindahl Index in Columns 1-2, and (2) using POLS in Columns 3-4. Panel A presents results for Hindi states. Panel B presents results for East Region. Hindi states are those states in which one of their official languages is Hindi. Individual controls encompass child sex, child age, household electricity connection, toilet availability, the household has a mobile, household type, the father attended school, and the mother attended school. Geographic Controls include latitude, longitude, and boundary district dummy. Demographic controls include the percentage of SCs population living in rural, percentage of STs population living in rural, percentage of the rural population, and percentage of the Muslim population. Standard errors are in parentheses and are clustered at the district level in all the specifications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6 Appendix

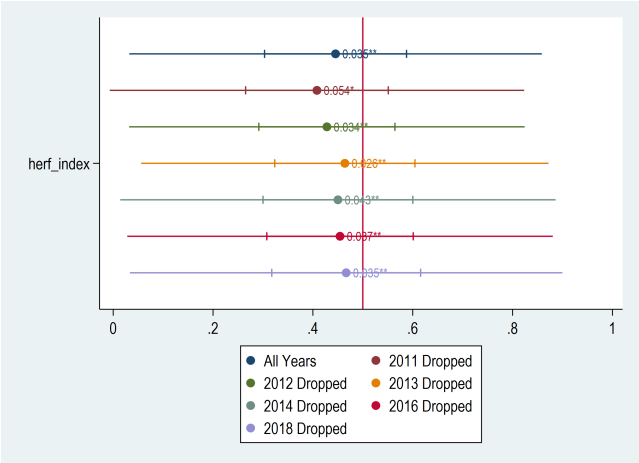

Figure 1: Coefficient Plot for Herfindahl Index

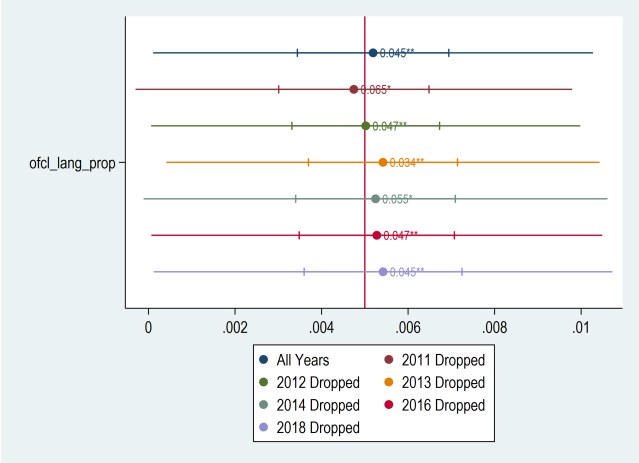

Figure 2: Coefficient Plot for POLS

Table A1: Heterogeneous Effect of Linguistic Minority Status For Major 18 States Read Score

	(1) First Stage	(2) Read Score	(3) First Stage	(4) Read Score
Panel A: Major 18 States				
Different Boundary	-.002323***		-.2065064***	
	(0.0005)		(0.0405)	
Herfindahl Index		.3602927*		
		(0.2182)		
POLS				.0040529
				(0.0025)
Observations	1862282	1862282	1862282	1862282
F-statistic		25.03666		25.97151
State Round Fixed Effect	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes
Geographic Controls	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes

Notes: This table depicts heterogeneous effects on the read scores of children aged 5-16 for major 18 states. Two specifications are employed: (1) utilizing the Herfindahl Index in Columns 1-2, and (2) using POLS in Columns 3-4. Panel A presents results for Hindi states. Panel B presents results for East Region. Individual controls encompass child sex, child age, household electricity connection, toilet availability, the household has a mobile, household type, the father attended school, and the mother attended school. Geographic Controls include latitude, longitude, and boundary district dummy. Demographic controls include the percentage of the SCs population living in rural, the percentage of the STs population living in rural, the percentage of the rural population, and the percentage of the Muslim population. Standard errors are in parentheses and are clustered at the district level in all the specifications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Heterogeneous Effect of Linguistic Minority Status For North, West, and South Regions on Read Score

	(1) First Stage	(2) Read Score	(3) First Stage	(4) Read Score
Panel A: North Region				
Different Boundary	-.0023505***		-.2682627**	
	(0.0008)		(0.1150)	
Herfindahl Index		.4113321		
		(0.3672)		
POLS				.003604
				(0.0034)
Observations	629384	629384	629384	629384
F-statistic		8.104685		5.440811
Panel B: West Region				
Different Boundary	.0003091		-.0746373	
	(0.0013)		(0.1147)	
Herfindahl Index		-.5070283		
		(4.5370)		
POLS				.0021
				(0.0172)
Observations	543112	543112	543112	543112
F-statistic		.0609945		.4236266
Panel C: South Region				
Different Boundary	-.0022928***		-.1436028***	
	(0.0005)		(0.0401)	
Herfindahl Index		-.0463363		
		(0.2233)		
POLS				-.0007398
				(0.0036)
Observations	308431	308431	308431	308431
F-statistic		19.66851		12.8137
State Round Fixed Effect	Yes	Yes	Yes	Yes
Individual Controls	Yes	Yes	Yes	Yes
Geographic Controls	Yes	Yes	Yes	Yes
Demographic Controls	Yes	Yes	Yes	Yes

Notes: This table depicts heterogeneous effects on the read scores of children aged 5-16 for North, West, and South regions. Two specifications are employed: (1) utilizing the Herfindahl Index in Columns 1-2, and (2) using POLS in Columns 3-4. Panel A presents results for Hindi states. Panel B presents results for East Region. Hindi states are those states in which one of their official languages is Hindi. Individual controls encompass child sex, child age, household electricity connection, toilet availability, the household has a mobile, household type, the father attended school, and the mother attended school. Geographic Controls include latitude, longitude, and boundary district dummy. Demographic controls include the percentage of the SCs population living in rural, the percentage of the STs population living in rural, the percentage of the rural population, and the percentage of the Muslim population. Standard errors are in parentheses and are clustered at the district level in all the specifications. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.