Ashoka University
Economics Discussion Paper 150

Is language a bridge or a barrier? Impact of linguistic distance on the health of women and children

July 2025

Anisha Sharma, Ashoka University
Advaith Jayakumar, Columbia University

https://ashoka.edu.in/economics-discussionpapers

Is language a bridge or a barrier? Impact of linguistic distance on the health of women and children

Anisha Sharma
Ashoka University
anisha.sharma@ashoka.edu.in

Advaith Jayakumar
Columbia University
aj3319@columbia.edu

July 29, 2025

Abstract

We measure the impact of a lack of familiarity with a dominant language on health outcomes and health-seeking behavior among women and children in India. We use language tree data from the Ethnologue to measure the linguistic distance between a person's mother tongue and the dominant language of the region they live in. We find evidence that increasing linguistic distance results in increased morbidity among women as well as reduced vaccine take-up for their children. Key mechanisms are reduced exposure to health information and decreased autonomy among women, making them less likely to be able to travel to a health clinic by themselves. Our results are robust to a number of alternative measures of linguistic distance, and suggest an added burden of being a migrant.

JEL Codes: O12, O15, I12, I14, I15

Keywords: development, health, linguistic distance

1 Introduction

Language plays a foundational role in shaping social interactions and enabling access to economic opportunities through employment, education, and access to public services. Yet, languages also differ widely in structure and composition, even within the same geographic areas, and these differences can create barriers to communication and integration (Bromham et al., 2015). These barriers are significant for migrants, in particular, as their native language often varies from the dominant language of the place in which they reside.

The potential returns to an investment in learning a new language are high (Chiswick, 2008; Ginsburgh and Weber, 2020). However, learning a new language is also costly, and the cost of acquiring a new language for an individual depends, in part, on how linguistically distinct the new language is from an individual's native tongue. For migrants, the more distinct their native tongue is from the dominant language of the region in which they reside, the greater the costs they face in achieving socioeconomic integration.

In this paper, we examine the consequences of linguistic barriers on access to healthcare, and, consequently, on health outcomes. Our study is set in India, a large developing country with considerable linguistic diversity across 22 official languages, and thousands of dialects in those languages. Proficiency in the local dominant language enables people to get access to the labour market, education, and healthcare (Laitin and Ramachandran, 2016). At the same time, there are a large number of internal migrants within the country. Census data from 2011 puts the count of rural-urban migrants at 51 million, but this is likely a significant underestimate, particularly if one accounts for temporary and seasonal migration as well (Tumbe, 2018; Singh et al., 2022). Given the extent of linguistic diversity, it is likely that a migrant's mother tongue is different from the dominant language of the region, imposing costs on migrants in terms of human capital outcomes.

Using data from two rounds of a nationally representative survey, the National Family Health Survey, and data on the distinctness between languages from Ethnologue (Lewis et al., 2014), we estimate the effect of the increase in the cost of acquiring a local dominant

language on observed health outcomes and health-seeking behaviour. To quantify the costs of acquiring a language, we use a measure of linguistic distance developed by Fearon (2003). The greater the linguistic distance between any two languages, the greater the cost for a native speaker of the language to learn the other. We find evidence that increasing linguistic distance between a woman's native language and the dominant language of the district she resides in results in poorer health outcomes for her. We focus on health outcomes that are likely to be responsive to receiving information on prevention and treatment from health services (for example, anaemia and high blood sugar). Our specification controls for a range of household and individual characteristics and district fixed effects, and is robust to the use of matching techniques so as to compare similar households that vary only in their linguistic distance from the dominant language of the region. This is particularly important for identification since our analysis effectively compares migrants to non-migrants.

We find that a one-unit increase in linguistic distance leads to a 0.9-1.3 percentage point increase in the probability of being anaemic or having high blood sugar levels. We also find evidence of reduced access to healthcare services for the children of these migrants: again, a one-unit increase in linguistic distance is associated with a reduced probability of receiving a vaccination by 3-6.9 percentage points. Our results are robust to using three different measures of linguistic distance. In terms of mechanisms, we find evidence of reduced health-seeking behavior by women, reduced exposure to public health information, as well as reduced autonomy of women to go outside the house and visit health facilities on their own – all of which increase with linguistic distance. Additionally, we explore heterogeneous treatment effects by household characteristics such as wealth and length of residence in a district. We find that increasing household wealth and an increase in the years for which a household is resident in a district moderates the negative effects of linguistic distance on the health outcomes of children. Finally, we consider heterogeneous treatment effects by the ethnolinguistic diversity of a district and whether the average income in the state is above or below the median for the country. We find that the negative effects of linguistic distance

are exacerbated in low-income states but moderated in districts with greater ethnolinguistic diversity.

This paper contributes to the literature on the adverse effects of linguistic barriers on human capital outcomes. A large literature has identified the negative effects of distance between the native language of a child and the effect of native language of instruction in schools on educational attainment (Laitin and Ramachandran, 2016; Chicoine, 2019; Laitin et al., 2019; Ginsburgh and Weber, 2020; Laitin and Ramachandran, 2022; Bernhofer and Tonin, 2022). For health outcomes, the evidence is somewhat mixed: some studies find a negative association between dominant language fluency and health outcomes (Ponce et al., 2006; Schachter et al., 2012; Pottie et al., 2008; Nguyen and Reardon, 2013), particularly for women (Guven and Islam, 2015; Dang, 2025) and their children (Black and Kunz, 2024), while others find no impact at all (Aoki and Santiago, 2018). Much of the existing literature examines variation in linguistic distance between immigrants' native languages and the dominant language spoken in their destination countries. In the context of a developing country, Laitin and Ramachandran (2016) use microdata from India to find that linguistic distance from the language of the government reduces awareness about health-improving behaviors such as the use of bed nets to protect against malaria, and awareness of AIDS. Gomes (2020) finds that in sub-Saharan Africa, increased linguistic distance from one's neighbours is associated with higher mortality and malnutrition for children. Our study also uses microdata from a large country, but we study a wider range of health outcomes and health-seeking behaviours, including anaemia, blood sugar, blood pressure, and child vaccine uptake.

We are also able to provide clear evidence on the mechanisms that likely explain our results. Prior research has shown that linguistic distance between healthcare providers and patients can lead to reduced trust and the quality of healthcare received (Street and Haidet, 2011). Language barriers also make it less likely that individuals, especially mothers, receive important information about healthcare (Narayan, 2013; Laitin and Ramachandran, 2016; Gomes, 2020). We also find support for the hypothesis that linguistic distance is correlated

with less exposure to media and information for mothers, and we show that this leads to lower health-seeking behaviour by these mothers. We also show that language barriers are correlated with the reduced autonomy of women and willingness to seek out health care outside the home, which compounds the difficulties in accessing information, especially in patriarchal societies like those in South Asia.

Section 2 describes the data. Section 3 discusses the empirical strategy. Section 4 presents the results. Section 5 concludes.

2 Data and descriptive statistics

2.1 Measuring linguistic distance

The cost to a native speaker of language A to learn another language B is directly related to the linguistic distance or distinctness between the two languages. For example, the linguistic distance between languages like Tamil and Kannada is low since they are similar in structure, reducing the cost for the speaker of one language to learn the other. On the other hand, the linguistic distance between Tamil and Nepali is high, since these are two very distinct languages, and the cost of learning one language by the native speaker of the other is high.

Measures of linguistic distance rely on 'language trees', which classify and group languages based on ancestry, origin, and structure, among other parameters. One such language tree is the Ethnologue (Lewis et al., 2014), which shows the relationship between different languages, specifically how languages evolved from common ancestors and split over time into different branches (see Figure 1 for an example). Several methods have been developed to compute the linguistic distance between any two languages in the language tree, based on how recently they diverged from one another. We use the Fearon (2003) method to compute linguistic distance based on data from the Ethnologue (Lewis et al., 2014). This measure is computed as follows:

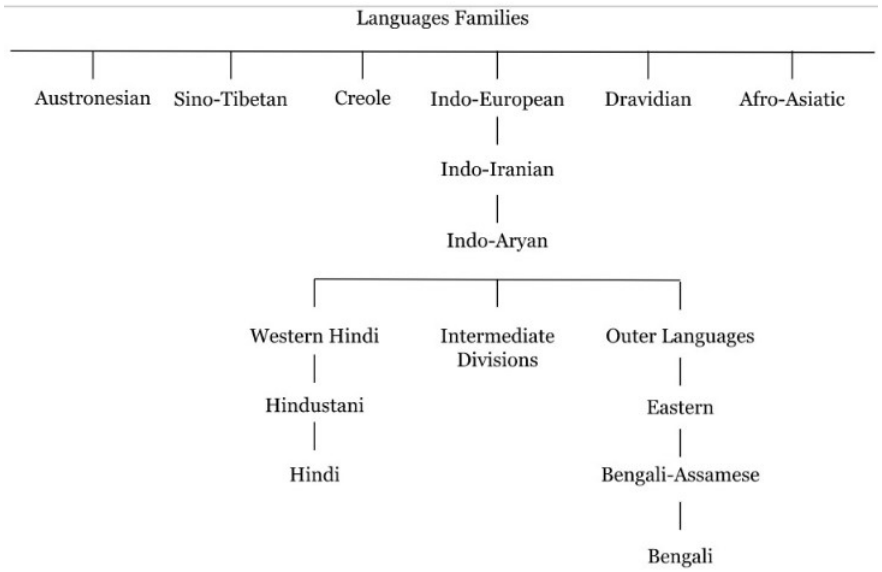

Figure 1: Language Families Tree from the Ethnologue

$$ld_{ij} = 1 - \left[\frac{\text{no. of common nodes between } i \text{ and } j}{\frac{\text{no. of nodes for } i + \text{no. of nodes for } j}{2}} \right]^{\lambda} \tag{1}$$

where ld_{ij} is the linguistic distance between between two languages, i and j. Here, if the languages belong to completely different language families, then the number of common nodes is 0, and the distance between the two languages is 1. The value of λ determines the relative distance between two languages that belong to the same family compared to two languages that belong to distinct language families. We follow Fearon (2003) in assuming that $\lambda = 0.5$, though we note that there is no strong theoretical justification for this choice.

We also use alternative measures of linguistic distance, such as those developed by Lewis et al. (2014) and used in studies such as Jain (2017), and the ASJP method developed by Wichmann et al. (2011). Further details on these measures are discussed in the appendix.

We next turn to describing the data and presenting descriptive statistics from our sample.

2.2 Data and descriptive statistics

To test the hypothesis that increasing linguistic distance is correlated with worsened health outcomes, we use pooled survey data from the two most recent rounds of the National Family

Health Survey conducted in 2015-16 (round 4) and 2019-21 (round 5). The pooled dataset includes 1,415,675 unique women between the ages of 15-49 years for whom data is available on some health outcomes. We also look at the relationship between linguistic distance and health investments in children, specifically under-5 immunization rates of approximately 130,000–370,000 children. The data also include a rich set of mother and household characteristics.

Linguistic Distance. The main source of variation in our data is linguistic distance (LD) between the respondent's mother tongue and the dominant language of the district. We identify the dominant language of the district from the 2011 Census as the language spoken by the most individuals in the district. 21 official languages are recorded for both the respondent's mother tongue and the dominant language of the district. Observations that list the mother tongue as "other" are dropped.[1] We, therefore, get a 21x21 matrix of pairwise linguistic distances for any two languages from among the 21 official languages, helping us characterise all possible combinations of mismatch between the dominant language and an individual's native language. 85% of the sampled women have an assigned LD of 0, i.e., there is no mismatch between their native language and the dominant language of the district. The remaining 15% of women in our sample have some mismatch between their native language and the dominant language of the region.[2] The size of this mismatch or distinctness varies from 0.13 (lowest LD between Hindi and Urdu) to 1 (highest LD between, for example, Malayalam and Marathi).[3]

Women's health. For the health outcomes of women, we consider the prevalence of anaemia, high blood pressure, and high blood sugar. These outcomes are chosen because they can be easily influenced by dietary choices and other environmental factors, and are

[1] This consists about 8% of the sample

[2] Additionally, in this sample, approximately 18% of women have a native language mismatch with their husbands.

[3] The NFHS only captures languages and not dialects. For example, the mother tongue of the North Indian belt is recorded as 'Hindi' even though there exist linguistic differences across the dialects of Hindi. The 2011 Census does capture dialects, but to harmonize the 2 datasets, we only consider languages, and not dialects, to compute the LD.

therefore amenable to change based on receiving health information from public and medical sources. The data includes a variable that captures a woman's haemoglobin levels: we create an indicator variable that takes the value 1 if the haemoglobin level is less than 12.0 gm/dL, indicating the prevalence of anaemia, and 0 otherwise.[4] Similarly, a woman is categorized as having 'high' blood pressure if her systolic reading is at least 140 mmHg and has 'high' blood sugar if her blood sugar level is at least 141 mg/dL.[5]

Investments in children. For children, we examine the immunization status of the child. There are two reasons for this. First, vaccine-preventable diseases can cause child stunting and long-term poor mental and physical health among adults (Nandi et al., 2020), and routine childhood vaccinations can significantly reduce the disease burden among children and improve their health outcomes. In India, free immunizations are provided to children under the age of 5 years, to protect against several diseases. However, the process of getting a child vaccinated requires communication with healthcare providers about follow-up dates, and health cards that document vaccine status are usually in the dominant language of the region. We therefore consider the impact of linguistic distance on routine vaccination for a number of vaccines for which information is available in the survey (DPT, Polio, Measles, Pentavalent, Rotavirus, Hepatitis B, Vitamin A1, and Vitamin A2 supplementation).

Table 1 shows differences between mothers with some level of mismatch between their native language and the dominant language of the district (i.e. with a linguistic distance ≥ 1) and mothers with no such mismatch (i.e. with a linguistic distance $= 0$) for a number of variables. In general, we see that women with a mismatch between their mother tongue and the dominant language spoken in their region have significantly worse individual health outcomes and lower uptake of immunizations for their children. There are also some statistically significant differences in socioeconomic characteristics across individuals with some mismatch and those with no mismatch that could affect health outcomes. For this reason, we control for these variables in the analysis that follows and also implement a matching

[4]This standard is based on that of the American Hematology Society and World Health Organization.
[5]These standards are based on NFHS-5 reports.

estimator to compare women who are very similar along all characteristics other than linguistic distance from the dominant language of the region. This is discussed in the next section.

3 Empirical strategy

We estimate the effect of linguistic distance between the dominant language and mother tongue on the incidence of poorer health outcomes among women and their children using the following specification:

$$Y_{idt} = \beta_0 + \beta_1 \text{LinguisticDistance}_{id} + \beta_2 X_{id} + \delta_d + \eta_t + \varepsilon_{idt} \tag{2}$$

Where Y_{idt} is the outcome variable for woman/child i currently residing in district d, surveyed in the year t. The independent variable is $LinguisticDistance_{id}$, which is the computed linguistic distance between the dominant language of the district d and the mother tongue of woman/child i. The identifying assumption here is that the linguistic distance between the mother tongue and the dominant language of the district is uncorrelated, conditional on controls, with other characteristics that could explain health. One could argue that there could be several socioeconomic and other demographic characteristics that are correlated to linguistic distance and health outcomes. To tackle this, we include X_{id}, which are covariates that capture socioeconomic and demographic characteristics, including age of the woman, her religion, caste, education, the household's wealth index, number of family members, number of children alive, sex of the household head and number of years spent in the current place of residence. We also include survey year and district fixed effects to account for spatial and intertemporal variation. For regressions of child outcomes, we also include the current age of the child, the sex of the child and birth order fixed effects as covariates. All standard errors are clustered at the household level since linguistic distance typically varies at the household level and our data has multiple female and child respondents from the same house.

To address concerns that migrant households are different from non-migrant households, even conditional on these controls, in ways that are correlated with health outcomes and health-seeking behaviour, we also provide results from estimation on a smaller sample of matched households, which we discuss in the next section.

4 Results

4.1 Main results

Impact on women's health. Our results on women's health outcomes are presented in Table 2. We consider three outcomes: whether the female respondent is anemic, whether she has high blood pressure, and whether she has high blood sugar.[6] We choose these outcomes since they are typically responsive to medical treatment that is easily available in public sector clinics. In panel A, the key explanatory variable is our measure of linguistic distance (LD). We find that the coefficient on LD is positive and significantly different from zero for anemia and high blood sugar, and positive but close to 0 for high blood pressure. A one-unit increase in linguistic distance between the mother tongue and the dominant language of the district is associated with a 1.3 pp increase in the probability of a woman being anaemic and a 0.9 pp increase in the probability of having high blood sugar. This implies that a one-standard-deviation increase in linguistic distance is associated with a 0.2 pp and 0.15 pp increase in the probability of a woman being anaemic and having high blood sugar, respectively.

In panel B, we use a different specification where we consider a binary measure of treatment instead of a continuous measure of linguistic distance: an indicator that takes the value 1 if the respondent's mother tongue is different from the dominant language of the district and 0 if it is the same. In other words, the variable 'Treatment' takes the value 1 for any

[6]The only other measured health outcomes captured in the survey relate to obesity. BMI and waist-to-hip ratio are both considered harder to interpret, since what constitutes a clinically adverse outcome can vary from person to person. As such, we exclude these measures from this analysis.

linguistic distance greater than 0 and 0 for a linguistic distance of 0 when comparing the respondent's mother tongue with the dominant language of the district she lives in. The coefficients on 'Treatment' are positive and significant for anemia, suggesting that a non-native speaker of the dominant language of a district is 1.9 pp more likely to be anemic, compared to a native speaker of the dominant language of a district. These results are consistent with the hypothesis that a growing linguistic barrier is leading to an increase in the incidence of poorer health outcomes among women. However, there is no impact on the incidence of high blood pressure and high blood sugar.

Given that migrant and non-migrant households do vary along multiple characteristics, we re-estimate the main specification in Panel B on a sample of matched households, where the matching is done using propensity scores. We predict the probability of a household being treated (having a linguistic distance between the native language and the dominant language of a district of greater than 0) as opposed to being untreated (where the native language is the same as the dominant language of the district) using the following variables: whether the household has a clean source of piped water, whether the household owns a fridge, whether the household has a flushed toilet facility, whether the household has clean cooking fuel, whether the household has a pucca floor material, pucca wall material and pucca roof material, and the number of women aged 15-49 years in the household. We then re-estimate our main results from Panel B on a matched sample, where individual observations are weighted by inverse of the propensity score from the matching estimation. Our results are similar to those obtained in the unmatched sample, although less precisely measured since the sample of matched households is significantly smaller than the main sample. The coefficient on treatment for anemia is positive and significant, and also higher than in the unmatched sample. A one unit increase in linguistic distance is associated with a 10.5 percentage point increase in the probability of a woman being anemic. The coefficient on high blood pressure is positive but not significantly different from 0. The effect on high blood sugar is negative but the coefficient is small. This suggests that differential selection

into migration is not driving our results.

Impact on child immunizations. We next turn to results on the vaccine status of children in Table 3. In panel A, coefficients on linguistic distance are negative and significantly different from zero for 6 out of 8 vaccines that we have information on, suggesting that as linguistic distance increases, vaccine take-up declines. The effect sizes vary from 3 pp to 6.9 pp declines in vaccine uptake for a one-unit increase in linguistic distance. Alternatively, a one-standard-deviation increase in linguistic distance is associated with a 0.5-1.1 pp decrease in vaccine uptake. When we consider results in panel B, where we use a binary variable for treatment, 7 out of 8 coefficients are negative, and 5 of the 7 are significantly different from 0. One coefficient (for rotavirus) is now positive and significantly different from 0, which is an anomalous result.

We test the robustness of our results by re-estimating them on a matched sample, and we find very similar results to those in panel B. 5 out of 8 coefficients are negative, and one is significantly different from 0, suggesting that greater linguistic distance is associated with poorer outcomes in terms of child vaccine uptake.

4.2 Mechanisms: Health-seeking behavior and female autonomy

We next turn to an examination of the specific channels through which linguistic distance from the dominant language might disadvantage people who speak a different language. We consider three possible channels. One possibility is that increased linguistic distance is correlated with reduced or inefficient engagement and communication with healthcare workers and services due to the language barrier. To examine this, we consider outcomes that capture engagement with the healthcare system. We consider whether a respondent has met with healthcare workers in the past three months, whether she has access to health insurance, whether she was informed about how to deal with the side effects of a medical procedure, sterilization, and whether they think the care they received after this medical procedure was adequate.

A second channel is exposure to information about public health through the media. Again, linguistic differences could make it harder for non-native speakers to access media offered in the local language. Since family planning programmes are key elements of public health outreach by the government, we focus on whether the respondent has been exposed to family planning messaging on any of the following mediums: radio, newspaper, or TV.

A third channel is differences in individual autonomy of women since families might impose greater restrictions on the mobility of women who are less familiar with the local language than women who speak the native language of a district. We consider multiple measures of autonomy.

Engagement with the healthcare process. We first examine the impact of increasing linguistic distance on a woman's engagement with healthcare workers and her experience with medical procedures. For this, we look at whether the woman has met with a specialised female healthcare worker such as an Auxiliary Nurse Midwife (ANM) or a lady health worker (LHV) in the last 3 months. For all women who have gone through the female sterilization surgical process, the NFHS collects data on whether the respondent was adequately informed about the side effects of the procedure and whether she received a high or low quality of care during and after the medical procedure. We also consider whether the woman is covered by health insurance and whether pregnant women have accessed healthcare. These measures can captures the respondent's willingness and ability to engage with healthcare workers as well as their experiences during medical treatment and post-treatment procedures.

The results are presented in Table 4 and paint a stark picture of reduced access to healthcare. We find that as linguistic distance increases by 1 unit, the probability of a woman having met a healthcare worker reduces by 3.5 pp (significant at the 1% level), and women are 0.6 pp less likely to report the quality of their healthcare as "good" (significant at 10%). Women are 2.5 pp less likely to be covered by health insurance (significant at 1%). Pregnant women are 2.9 pp less likely to have a health card (significant at 1%), 3.4 pp less likely to have visited an ANC worker during pregnancy (significant at 1%), 2 pp less likely

to have been told about pregnancy complications (significant at 5%) and 2.6 pp less likely to have been supplements during pregnancy (significant at 1%).

In short, our results suggest that women facing linguistic barriers engage less with the healthcare system and have less satisfying outcomes when they do.

Exposure to health-pertinent information through media. We next consider exposure to health-pertinent information, particularly through the media. In the context of India, (Laitin and Ramachandran, 2016) uses data from the NFHS-3, collected in 2005, to find that women are less likely to be aware of AIDS or use bed nets to reduce the incidence of malaria. They argue that both of these outcomes are measures of knowledge of health information and best practices, both of which can be mediated through exposure to health information through the media. We corroborate the results for these two outcomes in our sample of households pooled from the sample of NFHS-4 and NFHS-5 households and find similar effects (columns 4 and 5 of Table 5).

In addition, we explore some additional variables that directly measure exposure to health information through the media. Previous literature has found that increased familiarity with dominant languages improves the knowledge of health through exposure to the media (Ruiz et al., 1992). The variables we consider all relate to family planning. We estimate the impact of growing linguistic distance between the mother tongue and the dominant language of the region on whether a woman has heard of family planning methods through radio, television, and newspapers. Since the language of communication for these forms of media is primarily in the dominant language, women who do not speak this language well may be less likely to access this information. Table 5 presents these results; the estimated coefficients on linguistic distance are negative and significantly different from zero across all three media exposure outcomes. As linguistic distance increases by one unit, women are 6.1 pp less likely to have heard about family planning through radio, 1.3 pp less likely to have heard about family planning through newspapers, and 6.4 pp less likely to have heard about family planning from television. These effects are significant at the 1%, 10% and 1% level, respectively. This

provides some suggestive evidence that women facing language barriers may be less likely to access health-related information from the media.

Individual autonomy. The final channel to understand worsened health status for women facing language barriers is the reduced autonomy of women to visit healthcare facilities outside the home. Linguistic distance could be correlated with reduced female autonomy if a woman is restricted from leaving the house in an unfamiliar environment where she is less able to communicate with ease. We test this hypothesis by regressing different measures of women's autonomy on linguistic distance. We consider the following outcomes: whether the female respondent is allowed to leave the house alone to get medical help for herself, whether she can get medical help for herself at all, whether she is allowed to go to a medical facility by herself, whether she can leave the village by herself, and whether she can go to the market by herself. If women are only allowed to leave the house in the company of other men and women, then they may be less able or willing to seek out healthcare for themselves or their children.

The results are presented in Table 6 and indicate that as linguistic distance increases, all measures of autonomy decline. As linguistic distance increases by one unit, the probability of the woman being allowed to get medical help for herself alone or even with someone declines by 0.8 and 1.8 pp (Columns 1 and 2, significant at 5% and 1% respectively). More generally as well, women's mobility declines as linguistic distance increases: with a one unit increase in linguistic distance, women are 6.4 pp less likely to be allowed to go to a medical facility alone, 4.4 pp less likely to be allowed to go outside the village alone, and 6.9 pp less likely to be allowed to go to the market alone. All three results are significant at the 1 percent level.

Taken together, our results provide evidence that linguistic distance presents a barrier to accessing healthcare on multiple fronts. Women are less likely to leave home by themselves, they are less likely to engage with healthcare services, and they are less likely to receive health-relevant information through the media. These represent likely mechanisms that can explain the higher instances of morbidity and low vaccine uptake we see among respondents.

4.3 Heterogeneity

We next examine whether the effects of linguistic distance that we observed in the previous sections vary by household characteristics, such as household wealth and how long a household has been resident in a particular region. We also consider heterogeneity in treatment effects by characteristics of the district the respondents reside in, particularly whether the district is highly diverse according to a measure of ethnolinguistic diversity, and whether the state has relatively high state capacity.

Wealth Poorer families tend to rely more on public health officials, including government doctors and nurses, while people in higher wealth quintiles are more likely to access private healthcare. It is not theoretically clear whether public or private healthcare will be better equipped to deal with linguistic mismatches. If doctors speaking the same language as a patient or a patient's parent are more likely to be found in the private sector, then the adverse effects of linguistic distance should decline with wealth. Wealthier households may also be better able to compensate for the difficulty they face in accessing healthcare due to linguistic barriers, through access to higher-quality healthcare.

For this analysis, we use the five wealth quintiles that are derived from the wealth index in NFHS.We create a variable that takes the value 1 if the woman belongs to the top 2 quintiles, and 0 if not. In Table A1, Panel A, for morbidities of women, we find that the impact of linguistic distance on high blood pressure is moderated for rich households as compared to non-rich households. The coefficient on the interaction between linguistic distance and the indicator for whether a household is rich is significant at the 5% level. Similarly, for vaccine take-up, the coefficient on the interaction between rich and linguistic distance is positive and significantly different from 0 for 6 out of 8 outcomes, suggesting that linguistic distance is less costly for the rich Table A2. This suggests that the impact of linguistic distance falls more heavily on the poor than the rich. Public healthcare services, therefore, need to play a more active role in addressing linguistic gaps since the poor are more likely to use subsidised public services.

Years of residence. While languages are costly to learn, they are usually acquired over time. The longer a person stays in a particular region, the more likely their language skills are to improve, and thus the size of the barrier posed by language can be reduced over time. We hypothesize that the magnitude of the negative health effect will likely become smaller the longer a person has stayed in a particular region. To test this, we divide the sample based on the year of current residence for the respondent. We create an indicator variable for whether a woman has lived at least 5 years in her place of residencecreate three buckets: women who have spent less than five years in their current place of residence, women who have spent 5 to 9 years in their current place of residence, and women who have spent 10 or more years in their current place of residence. We interact the indicator variables for 5-9 years and 10 or more years with linguistic distance and examine the coefficients on the interaction terms. **Years of residence.** While languages are costly to learn, they are usually acquired over time. The longer a person stays in a particular region, the more likely their language skills are to improve, and thus the size of the barrier posed by language can be reduced over time. We hypothesize that the magnitude of the negative health effect will likely become smaller the longer a person has stayed in a particular region. To test this, we divide the sample based on the year of current residence for the respondent. We create an indicator variable for whether a woman has lived at least 5 years in her current place of residence and we regress health outcomes on the interaction between this variable and a measure of linguistic distance.

Table A1 presents the results for women and Table A2 for children. Surprisingly, we find that for women, the longer they have been present in a particular location, the worse their health outcomes are likely to be. The coefficients on the interaction between linguistic distance and the indicator for being resident for at least five years in the current place of residence is positive and significantly different from 0 for all three outcomes – anemia, high blood pressure, and high blood sugar. For vaccine uptake, the results go the other way, but are less strong. 2 out of 8 coefficients on the interaction between linguistic distance and the

indicator for residence of more than five years are positive and significantly different from 0. Thus, we do not have compelling evidence that the impact of linguistic distance reduces over time. In fact, for some morbidities, adverse outcomes even increase with the length of residence. This suggests that the barriers to healthcare access do not decline with time, even if that time brings some increasing familiarity with the local language. It could also suggest that unless timely medical interventions are made at the onset of a morbidity, improved communication over time will not reduce the prevalence of that condition.

Ethnolinguistic fractionalization. A large body of literature studies the relationship between ethnolinguistic diversity and socioeconomic development (Easterly and Levine, 1997; Alesina et al., 2003; Desmet et al., 2009, 2020). The nature of local linguistic diversity in the places where people live can also affect their ability to interact with local health services, and, hence, their health outcomes. For example, Kumar et al. (2020) finds that children facing a linguistic barrier do relatively poorly on language-dependent tasks, but only when they live in ethnically segregated communities, compared to ethnically mixed communities. To more accurately understand the role of other linguistic groups in possibly mitigating or worsening the language barrier between one's mother tongue and the dominant language of the region, we look at variation in treatment effects by an index of how linguistically diverse a region is. Linguistic diversity is typically measured using an index of ethnolinguistic fractionalization, or ELF. The standard ELF is constructed on the basis of the Herfindahl-Hirschman index and captures the diversity of ethnicities and languages in a population (Easterly and Levine, 1997). It is calculated as

$$ELF = 1 - \sum_{i}^{n} s_i^2$$

where s_i is the share of the population that belongs to linguistic group i and N is the total number of linguistic groups. The ELF measure captures the probability of two randomly selected individuals in a state/district population belonging to two distinct linguistic groups. Thus, the ELF index takes values between 0 and 1 with a value of 0 implying no linguistic diversity at all and 1 indicating that every inhabitant of the society belongs to a distinct

linguistic group. For our analysis, we compute the ELF of each district and we create an indicator variable which takes the value 1 for all respondents living in districts with an above-median level of the ELF index, and 0 otherwise. Table A1 for women and Table A2 show that, in general, individuals from more linguistically diverse districts exhibit better health outcomes with increasing language distance than individuals in less linguistically diverse districts. In other words, the effect of linguistic distance on women's health is moderated in highly fractionalized districts for anemia: the coefficient on the interaction between linguistic distance and the indicator for whether the respondent's district has an above-median level of ELF is negative and significantly different from 0 at the 1% level (Table A1). Similarly, the adverse impact of linguistic distance on children in high ELF districts is moderated compared to the impact on children in low ELF districts (Table A2). We see that the coefficient on the interaction term is positive for all 8 outcomes and statistically significantly different from 0 for 4 out of 8 outcomes. One reason for this could be that districts with higher linguistic diversity are better equipped to deal with language barriers in the healthcare system, leading to better health outcomes among children.

State capacity. Finally, we look at variation in the impact of linguistic distance on health outcomes by whether the respondent lives in a state with high or low state capacity. We proxy state capacity by state per capita income and create an indicator variable for low-income state, which takes the value 1 for all respondents living in states with a below-median level of per capita income (Bihar, Uttar Pradesh, Madhya Pradesh, Rajasthan, Jharkhand, Odisha, Chattisgarh, West Bengal) and 0 for all respondents in states with an above-median level of per capita income (Andhra Pradesh, Arunachal Pradesh, Assam, Goa, Gujarat, Haryana, Himachal Pradesh, Jammu and Kashmir, Karnataka, Kerala, Maharashtra, Manipur, Meghalaya, Mizoram, Nagaland, Punjab, Sikkim, Tamil Nadu, Telangana, Tripura, Uttarakhand and all union territories). The results are presented in Table A1. We find that the impact of linguistic distance on women's health is unaffected by the state in which a woman is located. However, children's vaccine uptake is significantly lower as linguistic

distance increases in low-income states for 3 out of 8 vaccines. In sum, states with higher incomes and higher capacity to manage public services are able to deliver some types of healthcare better than other states to people who face linguistic barriers.

4.4 Further results

Other measures of distinctness. As we discuss previously, there are different ways to measure distinctness between languages. Our main results rely on the measure of linguistic distance developed by Fearon (2003). In this section, we confirm that our results are robust to alternative definitions and measures of the size of the language barrier. We use methods developed by Lewis et al. (2014) and used in Jain (2017), and the Automated Similarity Judgment Program (ASJP) method developed by Wichmann et al. (2011). For our sample, given all the combinations of language mismatch, the linguistic distance measured by Lewis et al. (2014) varies between 3-16 for families who speak a different language from the dominant language of their district (a score of 0 indicates no mismatch). Similarly, by the ASJP method, the linguistic distance varies between 34.84-104.16 (a score of 0 indicates no mismatch). Given the variation in the measure of linguistic distance, the size of the estimated effects will be quantitatively different but the sign and direction of the coefficients on linguistic distance for different outcomes should be consistent with our main results.

We present results on female morbidities in Tables A3. For the three main outcome variables of anemia, high blood pressure and high blood sugar, the sign on linguistic distance is positive and significantly different from zero for most outcomes. For the ASJP method, the positive coefficient on linguistic distance for anemia is positive but not significantly different from 0. Similarly for vaccine uptake among the children of respondents, Table A4 presents results that are consistent with the main findings, showing that vaccine uptake declines with linguistic distance.

5 Conclusion

This paper contributes to the literature on the adverse effects of linguistic distance on human capital accumulation, specifically through access to the healthcare system. Using data from the National Family Health Survey, we find that increased linguistic distance between a woman's mother tongue and the dominant language in her place of residence is associated with poorer health outcomes for mothers and their children, including higher rates of individual morbidities and lower take up of routine vaccines. Our results are robust to using three different measures of linguistic distance. The study identifies reduced health-seeking behavior, reduced exposure to health information obtained through the media, and diminished autonomy among women as mechanisms driving these effects. We also find that the adverse effects of linguistic distance can be moderated by household-level characteristics like wealth and by district and state-level characteristics, such as whether the district is ethnically diverse and whether the state has a relatively high capacity to deliver healthcare.

Language acquisition is a critical channel for social, economic, and cultural mobility. The design of inclusive policies that allow linguistically diverse groups of people to more fully engage with the state can play a pivotal role in reducing disparities in health and other socio-economic indicators of development. Our results suggest policy-makers should make greater effort to include linguistically diverse groups in their programmes for healthcare outreach to improve their health outcomes. Alternatively, enabling individuals to more easily learn dominant languages could also reduce the costs of acquiring new languages and mitigate the adverse effects of linguistic distance. This is especially important in linguistically diverse countries like India with large flows of internal migrants.

References

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg, "Fractionalization," *Journal of Economic Growth*, 2003, *8* (2), 155–194.

Aoki, Yu and Lualhati Santiago, "Speak better, do better? Education and health of migrants in the UK," *Labour Economics*, 2018, *52*, 1–17.

Bernhofer, Juliana and Mirco Tonin, "The effect of the language of instruction on academic performance," *Labour Economics*, 2022, *78*, 102218.

Black, Nicole and Johannes S Kunz, "The intergenerational effects of language proficiency on child health outcomes: Evidence from survey-and Census-matched health care records," *Journal of Economic Behavior & Organization*, 2024, *225*, 136–152.

Bromham, Lindell, Xia Hua, Thomas G Fitzpatrick, and Simon J Greenhill, "Rate of language evolution is affected by population size," *Proceedings of the National Academy of Sciences*, 2015, *112* (7), 2097–2102.

Chicoine, Luke, "Schooling with learning: The effect of free primary education and mother tongue instruction reforms in Ethiopia," *Economics of Education Review*, 2019, *69*, 94–107.

Chiswick, Barry R., "The Economics of Language: An Introduction and Overview," Technical Report, Working Paper 2008.

Dang, Thang, "Language training, refugees' healthcare integration, and the next generation's health," *Journal of Development Economics*, 2025, p. 103470.

Desmet, Klaus, Joseph Flavian Gomes, and Ignacio Ortuño-Ortín, "The geography of linguistic diversity and the provision of public goods," *Journal of Development Economics*, 2020, *143*, 102384.

_ , Shlomo Weber, and Ignacio Ortuño-Ortín, "Linguistic diversity and redistribution," *Journal of the European Economic Association*, 2009, *7* (6), 1291–1318.

Easterly, William and Ross Levine, "Africa's growth tragedy: Policies and ethnic divisions," *The Quarterly Journal of Economics*, 1997, pp. 1203–1250.

Fearon, James D, "Ethnic and cultural diversity by country," *Journal of Economic Growth*, 2003, *8*, 195–222.

Ginsburgh, Victor and Shlomo Weber, "The economics of language," *Journal of Economic Literature*, 2020, *58* (2), 348–404.

Gomes, Joseph Flavian, "The health costs of ethnic distance: Evidence from Sub-Saharan Africa," *Journal of Economic Growth*, 2020, *25* (2), 195–226.

Guven, Cahit and Asadul Islam, "Age at migration, language proficiency, and socioeconomic outcomes: Evidence from Australia," *Demography*, 2015, *52* (2), 513–542.

Jain, Tarun, "Common tongue: The impact of language on educational outcomes," *The Journal of Economic History*, 2017, *77* (2), 473–510.

Kumar, Hemanshu, Rohini Somanathan, Mahima Vasishth et al., "Language and Learning in Ethically Mixed Communities: A Study of School Children in an Indian Village," Technical Report 2020.

Laitin, David D and Rajesh Ramachandran, "Language policy and human development," *American Political Science Review*, 2016, *110* (3), 457–480.

__ **and** __ , "Linguistic diversity, official language choice and human capital," *Journal of Development Economics*, 2022, *156*, 102811.

__ **,** __ **, and Stephen L Walter**, "The legacy of colonial language policies and their impact on student learning: Evidence from an experimental program in Cameroon," *Economic Development and Cultural Change*, 2019, *68* (1), 239–272.

Lewis, M Paul, G Simon, and P Fennig, *Ethnologue: Languages of the world*, Dallas, Texas: SIL International, 2014.

Nandi, Arindam, Santosh Kumar, Anita Shet, David E. Bloom, and Ramanan Laxminarayan, "Childhood vaccinations and adult schooling attainment: Long-term evidence from India's Universal Immunization Programme," *Social Science & Medicine*, 2020, *250*, 112885.

Narayan, Lalit, "Addressing language barriers to healthcare in India," *National Med J India*, 2013, *26* (4), 236–8.

Nguyen, Duy and Leigh J. Reardon, "The role of race and English proficiency on the health of older immigrants," *Social Work in Health Care*, 2013, *52* (6), 599–617.

Petroni, Filippo and Maurizio Serva, "Measures of lexical distance between languages," *Physica A: Statistical Mechanics and its Applications*, 2010, *389* (11), 2280–2283.

Ponce, Ninez A., Ron D. Hays, and William E. Cunningham, "Linguistic disparities in health care access and health status among older adults," *Journal of General Internal Medicine*, 2006, *21* (7), 786–791.

Pottie, Kevin, Edward Ng, Denise Spitzer, Alia Mohammed, and Richard Glazier, "Language proficiency, gender and self-reported health: An analysis of the first two waves of the longitudinal survey of immigrants to Canada," *Canadian Journal of Public Health*, 2008, *99*, 505–510.

Ruiz, Monica S., Gary Marks, and Jean L. Richardson, "Language acculturation and screening practices of elderly Hispanic women: The role of exposure to health-related information from the media," *Journal of Aging and Health*, 1992, *4* (2), 268–281.

Schachter, Ariela, Rachel T Kimbro, and Bridget K Gorman, "Language proficiency and health status: Are bilingual immigrants healthier?," *Journal of health and social behavior*, 2012, *53* (1), 124–145.

Singh, Rajni, Navneet Manchanda, and Rakesh Mishra, "Internal student migration in India: Impact of the COVID-19 crisis," *Asian and Pacific Migration Journal*, 2022, *31* (4), 454–477.

Street, Richard L. and Paul Haidet, "How well do doctors know their patients? Factors affecting physician understanding of patients' health beliefs," *Journal of General Internal Medicine*, 2011, *26* (1), 21–27.

Tumbe, Chinmay, *India moving: A history of migration*, Penguin Random House India Private Limited, 2018.

Wichmann, Søren, Taraka Rama, and Eric W. Holman, "Phonological diversity, word length, and population sizes across languages: The ASJP evidence," *Linguistic Typology*, 2011, *15* (2), 177–197.

Table 1: Balance Tables

	Treatment (Language Mismatch)	Control (No Language Mismatch)	Difference (Treatment - Control)	SE	N
Anemia	0.544	0.501	-0.043***	(0.001)	1373270
High Blood Pressure	0.040	0.051	0.011***	(0.000)	1400449
High Blood Sugar	0.059	0.054	-0.006***	(0.001)	1369515
Polio	0.472	0.553	-0.080***	(0.002)	378706
Hepatitis B	0.373	0.431	-0.059***	(0.002)	374194
Vitamin A1	0.613	0.667	-0.054***	(0.002)	375548
Vitamin A2	0.180	0.204	-0.024***	(0.002)	375073
DPT	0.664	0.721	-0.057***	(0.002)	377484
Measles	0.276	0.306	-0.029***	(0.003)	131325
Pentavalent	0.658	0.705	-0.048***	(0.003)	131331
Rotavirus	0.273	0.330	-0.057***	(0.003)	130814
No. of Household Members	5.637	5.293	0.344***	(0.005)	1523692
Total No. of Children Ever Born	1.790	1.799	-0.008**	(0.004)	1523692
Scheduled Caste	0.204	0.104	0.100***	(0.001)	1523692
Scheduled Tribe	0.121	0.481	-0.360***	(0.001)	1523692
OBC	0.427	0.200	0.227***	(0.001)	1523692
Respondent's Current Age	30.237	30.355	-0.118***	(0.021)	1523692
Poorest wealth quintile	0.191	0.242	-0.051***	(0.001)	1523692
Poorer wealth quintile	0.210	0.208	0.002***	(0.001)	1523692
Middle wealth quintile	0.211	0.194	0.017***	(0.001)	1523692
Richer wealth quintile	0.204	0.186	0.019***	(0.001)	1523692
Richest wealth quintile	0.183	0.170	0.013***	(0.001)	1523692
Hindu	0.796	0.533	0.263***	(0.001)	1523692
Muslim	0.127	0.143	-0.016***	(0.001)	1523692
Christian	0.041	0.221	-0.180***	(0.001)	1523692
No Education	0.244	0.260	-0.016***	(0.001)	1523692
Primary School level	0.121	0.121	0.000	(0.001)	1523692
Secondary School level	0.499	0.514	-0.015***	(0.001)	1523692
High School Level	0.136	0.105	0.031***	(0.001)	1523692
Male Household Head	0.853	0.858	-0.005***	(0.001)	1523692
Urban/Rural	1.725	1.759	-0.034***	(0.001)	1523692
Years Lived in Place of Residence	15.901	16.198	-0.298***	(0.023)	1523692
Round of Survey	4.540	4.548	-0.009***	(0.001)	1523692

Source. NFHS-4 and NFHS-5.

Table 2: Impact on Morbidities of Women

	(1)	(2)	(3)
	Anemia	High Blood Pressure	High Blood Sugar
Panel A: Linguistic distance (LD) as a continuous variable			
Linguistic Distance	0.013***	0.001	0.009***
	(0.005)	(0.002)	(0.002)
Observations	1,148,023	1,109,251	1,144,969
Mean	0.545	0.037	0.059
Panel B: Linguistic distance as a binary variable			
Treatment (Language Mismatch)	0.019***	0.001	0.001
	(0.002)	(0.001)	(0.001)
Observations	1,311,577	1,256,815	1,308,047
Mean	0.537	0.039	0.058
Panel C: Matched sample			
Treatment (Language Mismatch)	0.105**	0.016	-0.006
	(0.043)	(0.011)	(0.015)
Observations	185,457	169,617	184,955
Mean	0.496	0.048	0.054

Notes. The dependent variable is a binary indicator for whether the respondent is anemic, has high blood pressure, or has high blood sugar. Panel A presents results where the linguistic distance of the native language of the respondent from the dominant language of the district is measured as a continuous variable. Panel B presents results where Treatment takes the value of 1 if the linguistic distance is greater than 0, and 0 otherwise. Panel C presents results on the matched sample. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, the number of years in current place of residence and year of survey and district fixed effects, Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Source. NFHS-4 and NFHS-5.

Table 3: Impact on Child Immunisations

	(1) Polio	(2) Hepatitis B	(3) Vitamin A1	(4) Vitamin A2	(5) DPT	(6) Measles	(7) Pentavalent	(8) Rotavirus
Panel A: Linguistic distance (LD) as a continuous variable								
Linguistic Distance	-0.038***	-0.031***	-0.056***	-0.012	-0.030***	-0.054***	-0.069***	0.001
	(0.009)	(0.009)	(0.008)	(0.008)	(0.008)	(0.011)	(0.013)	(0.010)
Observations	315,120	311,652	312,555	312,164	314,334	109,106	109,122	108,649
Mean	0.553	0.432	0.667	0.203	0.721	0.307	0.705	0.333
Panel B: Linguistic distance as a binary variable								
Treatment								
(Language Mismatch)	-0.011***	-0.016***	-0.032***	-0.007**	-0.003	-0.003	-0.020***	0.018***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.006)	(0.006)	(0.006)
Observations	363,197	358,897	360,167	359,717	362,072	125,357	125,361	124,862
Mean	0.539	0.421	0.656	0.198	0.711	0.303	0.699	0.324
Panel C: Matched sample								
Treatment								
(Language Mismatch)	-0.040	-0.116***	0.036	-0.012	-0.003	0.038	-0.031	0.017
	(0.046)	(0.041)	(0.041)	(0.037)	(0.037)	(0.040)	(0.039)	(0.024)
Observations	53,733	52,904	53,241	53,164	53,408	16,652	16,651	16,616
Mean	0.486	0.383	0.623	0.187	0.672	0.291	0.673	0.266

Notes. The dependent variable is a binary indicator for whether the child has received all doses of a particular immunization. Panel A presents results where the linguistic distance of the native language of the respondent from the dominant language of the district is measured as a continuous variable. Panel B presents results where Treatment takes the value of 1 if the linguistic distance is greater than 0, and 0 otherwise. Panel C presents results on the matched sample. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, whether the child is a female, the number of years in the current place of residence and year of survey, birth-order, the child's birth year and district fixed effects, Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.
Source. NFHS-4 and NFHS-5.

Table 4: Mechanisms: Information about Public Health and Health-Seeking Behavior

	(1) Has Health Card	(2) Received ANC for Pregnancy	(3) Told about pregnancy complications	(4) Given supplements during Pregnancy
Linguistic Distance	-0.029***	-0.034***	-0.020**	-0.026***
	(0.009)	(0.006)	(0.008)	(0.007)
Observations	109,825	151,206	270,191	303,309
Mean	0.948	0.945	0.706	0.827
	(5) Met with an ANM/LHV Worker	(6) Health Insurance	(7) Informed about Side Effects	(8) Quality of Care
Linguistic Distance	-0.035***	-0.025***	0.003	-0.006*
	(0.004)	(0.005)	(0.007)	(0.003)
Observations	1,190,303	1,190,311	356,157	316,860
Mean	0.170	0.250	0.809	0.965

Notes. In Column 1, the dependent variable is a binary indicator for whether the woman has a health card or not. In Column 2, the respondent is asked whether she received antenatal care during her pregnancy. The dependent variable in Column 3 is a binary indicator for whether the woman was told by a healthcare worker about pregnancy complications. In Column 4, the dependent variable is also a binary for whether the woman was given supplements during her last pregnancy. In the bottom panel, Column 5, the dependent variable is a binary indicator for whether the woman has met with an ANM/LHV worker. Column 6 describes whether the woman is covered by health insurance. In Column 7, the dependent variable is a binary indicator for whether the respondent has been informed about how to deal with the side effects associated with a medical procedure (sterilization). In Column 8, the dependent variable is a binary indicator for whether the quality of care received after a medical procedure (sterilization in this case) was satisfactory or not. All estimations in both panels control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, the number of years in the current place of residence and year of survey and district fixed effects. Additionally, in the top panel, the estimations also include birth-order and child birth year fixed effects. Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Source. NFHS-4 and NFHS-5.

Table 5: Mechanisms: Information about Public Health

	(1) Source of FP: Radio	(2) Source of FP: Newspaper	(3) Source of FP: TV	(4) Respondent Slept Under Bed Net	(5) Heard of Aids
Linguistic Distance	-0.061***	-0.013*	-0.064***	-0.027***	-0.029***
	(0.005)	(0.007)	(0.006)	(0.005)	(0.010)
Observations	600,952	600,952	600,952	600,960	90,106
Mean	0.151	0.590	0.352	0.210	0.877

Notes. The dependent variables are a binary indicator of whether the Radio is a source of information about family planning in Column 1. In Column 2, the variable is a binary indicator for whether the newspaper is a source of information about family planning. In Column 3, the dependent variable is a binary indicator of whether television is a source of information about family planning. In Column 4, the dependent variable is a binary indication of whether the respondent slept under a bed net. In column 5, the dependent variable is a binary for whether the respondent has ever heard of aids. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, the number of years in the current place of residence and year of survey and district fixed effects, Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.
Source. NFHS-4 and NFHS-5.

Table 6: Mechanisms: Autonomy of Women

	(1) Allowed to go Alone and get Medical Help for Self	(2) Allowed to get Medical Help for Self	(3) Allowed to go to a Medical Facility Alone	(4) Allowed to go Outside this Village Alone	(5) Allowed to go Market Alone
Linguistic Distance	-0.008**	-0.018***	-0.064***	-0.044***	-0.069***
	(0.004)	(0.004)	(0.011)	(0.011)	(0.011)
Observations	1,190,311	1,190,311	191,238	191,238	191,238
Mean	0.186	0.153	0.495	0.545	0.481

Notes. The dependent variables are a binary indicator of whether going alone to get medical help for oneself is a problem or not in Column 1. In Column 2, the variable is a binary indicator for whether the respondent is allowed to get medical help for self. In Column 3, the dependent variable is a binary indicator for whether the respondent is allowed to go to a medical facility alone. In Column 4, the respondent is asked whether they are allowed to go to the market alone. In Column 5, the respondent is asked whether they are allowed to go outside their village alone. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, the number of years in the current place of residence and year of survey and district fixed effects, Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Source. NFHS-4 and NFHS-5.

6 Appendix

6.1 Alternative measures of linguistic distance

In addition to the method proposed by Fearon (2003), we show our results are robust to using alternative measures proposed by Jain (2017) and Wichmann et al. (2011). These are discussed below.

The (Lewis et al., 2014) method used by (Jain, 2017) method computes the number of nodes between each language pair on the language tree. We trace the nodes from the end node of one language to the common nodes that the language pair shares till the end node of the other language. In this approach, the more nodes between the two languages (i.e. the farther the languages are on the tree), the more distinct they are in terms of structure, grammar, and other linguistic parameters. Thus, to calculate the linguistic distance between two languages, we count the total number of nodes between them, including the end nodes. As an example, in Figure 1, the distance between Hindi and Bengali can be traced from Hindi (1 node) → Hindustani (2) → Western Hindi (3) → Indo-Aryan (4) → Outer Languages (5) → Eastern (6) → Bengali-Assamese (7) → Bengali (8). Thus, the resulting linguistic distance is 8.

Another method of measuring distinctness between languages uses the concept of lexical or lexicostatistical distance between languages. This is based on identifying similarities between common roots of words and shared vocabularies between languages. The measure of distance is based on the percentage of shared cognates between two languages (cognate words are words in any two languages that share similar meaning, spelling, and pronunciation). An automated method proposed by Petroni and Serva (2010) to measure lexical distance uses a normalization of a Levenshtein distance between two words – the minimum number of insertions, deletions, or substitutions of a single character needed to transform one word into the other. Wichmann et al. (2011) develop an ASJP (Automated Similarity Judgment Program) software to calculate this distance. We use the same measure to calculate the

distance between languages in our sample.

Table A1: Women: Heterogeneity

	(1)	(2)	(3)
		High	High
	Anemia	Blood Pressure	Blood Sugar
Panel A: Household Wealth			
LD x Rich	-0.008	-0.006**	0.001
	(0.006)	(0.002)	(0.003)
LD	0.016***	0.003	0.008***
	(0.005)	(0.002)	(0.003)
Observations	1,148,023	1,109,251	1,144,969
Mean	0.545	0.037	0.059
Panel B: Years in Residence			
LD X 5 years	0.012*	0.006**	0.007**
	(0.007)	(0.002)	(0.003)
LD	0.003	-0.004	0.004
	(0.007)	(0.002)	(0.003)
Observations	1,148,023	1,109,251	1,144,969
Mean	0.545	0.037	0.059
Panel C: High Vs Low ELF Districts			
LD X High ELF	-0.061***	0.004	0.003
	(0.023)	(0.008)	(0.011)
LD	0.005	0.001	0.011***
	(0.005)	(0.002)	(0.003)
Observations	1,148,023	1,109,251	1,144,969
Mean	0.545	0.037	0.059
Panel D: Low Vs High Income States			
LD X Low Income State	-0.002	0.005	-0.010
	(0.015)	(0.006)	(0.007)
LD	0.016***	0.000	0.009***
	(0.005)	(0.002)	(0.002)
Observations	1,148,023	1,109,251	1,144,969
Mean	0.545	0.037	0.059

Notes. The dependent variable is a binary indicator of whether the respondent is anemic, has high blood pressure, or has high blood sugar. In Panel A, the dummy variable 'Rich' takes the value 1 when the respondents belong to the top two quintiles and are equal to 0 when the women belong to the bottom three quintiles. In Panel B, '5 years' is an indicator that takes the value 1 when a woman has resided in a place for more than 5 years and 0 otherwise. In Panel C, the indicator variable 'High ELF' takes the value 1 when the woman resides in a district with an above-median ELF, and 0 otherwise. In Panel D, 'Low Income State' takes the value 1 for a woman residing in a state with below-median state per capita income, and 0 otherwise. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, the number of years in the current place of residence and year of survey and district fixed effects. Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Source. NFHS-4 and NFHS-5.

Table A2: Children: Heterogeneity

	(1) Polio	(2) Hepatitis B	(3) Vitamin A1	(4) Vitamin A2	(5) DPT	(6) Measles	(7) Pentavalent	(8) Rotavirus
Panel A: Household Wealth								
LD X Rich	0.021*	0.020*	0.027**	0.025**	0.042***	0.021	0.050***	0.007
	(0.012)	(0.011)	(0.011)	(0.010)	(0.010)	(0.017)	(0.019)	(0.015)
LD	-0.046***	-0.039***	-0.067***	-0.022***	-0.048***	-0.062***	-0.088***	-0.001
	(0.010)	(0.010)	(0.010)	(0.008)	(0.009)	(0.013)	(0.015)	(0.010)
Panel B: Years in Residence								
LD X 5 Years	0.019*	0.020*	0.001	0.003	0.004	0.006	0.007	-0.002
	(0.011)	(0.010)	(0.011)	(0.009)	(0.010)	(0.016)	(0.018)	(0.012)
LD	-0.046***	-0.040***	-0.056***	-0.013	-0.032***	-0.057***	-0.072***	0.003
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.014)	(0.016)	(0.012)
Panel C: High Vs Low ELF Districts								
LD X High ELF	0.167***	0.135***	0.176***	0.002	0.156***	0.027	0.119	0.006
	(0.046)	(0.040)	(0.053)	(0.040)	(0.053)	(0.063)	(0.092)	(0.042)
LD	-0.040***	-0.039***	-0.056***	-0.011	-0.027***	-0.064***	-0.053***	-0.001
	(0.010)	(0.010)	(0.009)	(0.008)	(0.009)	(0.013)	(0.015)	(0.012)
Panel D: Low Vs High Income States								
LD X Low Income State	-0.009	-0.061**	-0.022	-0.010	-0.007	-0.083**	-0.056	-0.114***
	(0.028)	(0.027)	(0.027)	(0.021)	(0.024)	(0.039)	(0.039)	(0.038)
LD	-0.033***	-0.022**	-0.057***	-0.014*	-0.030***	-0.041***	-0.058***	0.006
	(0.009)	(0.010)	(0.009)	(0.008)	(0.009)	(0.012)	(0.014)	(0.010)
Observations	315,120	311,652	312,555	312,164	314,334	109,106	109,122	108,649
Mean	0.553	0.432	0.667	0.203	0.721	0.307	0.705	0.333

Notes. The dependent variable is a binary indicator for whether the child has received all doses of a particular immunization. In Panel A, the dummy variable 'Rich' takes the value 1 when the respondents belong to the top two quintiles and are equal to 0 when the respondents belong to the bottom three quintiles. In Panel B, '5 years' is an indicator that takes the value 1 when a child's mother has resided in a place for more than 5 years and 0 otherwise. In Panel C, the indicator variable 'High ELF' takes the value 1 when the child resides in a district with an above-median ELF, and 0 otherwise. In Panel D, 'Low Income State' takes the value 1 for a child residing in a state with below-median state per capita income, and 0 otherwise. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, whether the child is a female, the number of years in the current place of residence and year of survey, birth-order, the child's birth year and district fixed effects, Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Source. NFHS-4 and NFHS-5.

Table A3: Other Linguistic Distance Measures on Morbidities

	(1)	(2)	(3)
	Anemia	High Blood Pressure	High Blood Sugar
Panel A: Linguistic distance (LD) using the Lewis(2014) method			
Linguistic Distance	0.687**	0.026	0.472***
	(0.327)	(0.135)	(0.160)
Panel B: Linguistic distance (LD) using the ASJP method			
Linguistic Distance	0.078**	0.000	0.060***
	(0.039)	(0.016)	(0.019)
Observations	1,147,900	1,109,129	1,144,846
Mean	0.545	0.037	0.059

Notes. The dependent variable is a binary indicator for whether the respondent is anemic, has high blood pressure, or has high blood sugar. In Panel A, 'Linguistic Distance' is computed using the Lewis(2014) method and is divided by 1000. In Panel B, 'Linguistic Distance' is computed using the ASJP method and is divided by 1000. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, the number of years in the current place of residence and year of survey and district fixed effects, Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.
Source. NFHS-4 and NFHS-5.

Table A4: Other Linguistic Distance Measures on Immunisations

	(1) Polio	(2) Hepatitis B	(3) Vitamin A1	(4) Vitamin A2	(5) DPT	(6) Measles	(7) Pentavalent	(8) Rotavirus
Panel A: LD using the Lewis(2014) method								
Linguistic Distance	-2.883***	-2.456***	-4.074***	-1.269***	-2.051***	-4.009***	-5.035***	0.075
	(0.566)	(0.561)	(0.539)	(0.462)	(0.499)	(0.845)	(0.880)	(0.817)
Panel B: Linguistic distance (LD) using the ASJP method								
Linguistic Distance	-0.353***	-0.295***	-0.497***	-0.150***	-0.243***	-0.453***	-0.625***	-0.014
	(0.067)	(0.066)	(0.064)	(0.055)	(0.059)	(0.100)	(0.104)	(0.097)
Observations	315,093	311,625	312,528	312,137	314,307	109,101	109,117	108,644
Mean	0.553	0.432	0.667	0.203	0.721	0.307	0.705	0.333

Notes. The dependent variable is a binary indicator for whether the child has received all doses of a particular immunization. In Panel A, 'Linguistic Distance' is computed using the Lewis(2014) method and is divided by 1000. In Panel B, 'Linguistic Distance' is computed using the ASJP method and divided by 1000. All estimations control for the age of the respondent, the highest level of education completed, number of children she has, number of household members, caste, religion, wealth quintile, whether the household head is male, whether the household is located in a rural area, whether the child is a female, the number of years in the current place of residence and year of survey, birth-order, the child's birth year and district fixed effects, Standard errors clustered at the household level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.
Source. NFHS-4 and NFHS-5.