Online Appendix: Not for Publication

Language Barriers, Technology Adoption and Productivity:
Evidence from Agriculture in India

A Cell-level Measure of Technology Adoption and Agricultural Yields

Data on technology adoption is sourced from the Agricultural Input Survey (AIS), conducted at five-year intervals by the Ministry of Agriculture in coincidence with the Agricultural Census to collect information on input use by Indian farmers. In the survey, all operational holdings from a randomly selected 7% sample of all villages in a subdistrict are interviewed about their input use.[1] The AIS reports information on land farmed with each technology – or combination of technologies – at the district-crop level.

We construct the share of land farmed with a given agricultural technology k in a given cell i using the following neutral assignment rule:

$$\left(\frac{Area^k}{Area}\right)_{idt} = \sum_{c \in C_i} \left[\left(\frac{Area_{idc,t=2000}}{Area_{id,t=2000}}\right) \times \left(\frac{Area^k}{Area}\right)_{dct} \right] \tag{1}$$

The first element in the summation is the share of land farmed with crop c in cell i, which is observed at cell level in the FAO-GAEZ dataset and captures the initial allocation of land across crops in a given cell in the baseline year 2000.[2] The second element in the summation is the share of land farmed with technology k in district d among the land farmed with crop c. This variable captures the rate of technology adoption for a given crop in a given district and varies over time. Thus, the product of these two elements gives us an estimate of the share of land in cell i that is farmed under technology k and crop c. Summing across the set of crops farmed in cell i (C_i), we obtain an estimate of the share of land farmed with a given technology in a given cell.[3]

Similarly, we construct the measure of agricultural yield in a given cell using the following rule:

$$yield_{idt} = \sum_{c \in C_i} \left[\left(\frac{Area_{idc,t=2000}}{Area_{id,t=2000}}\right) \times \left(\frac{yield_{dct}}{yield_{dc}}\right) \right] \tag{2}$$

where $yield_{dct}$ is the production of the crop c per unit area cropped in district d in the year t. We normalize the yield across crops by dividing the district-crop-time yield $yield_{dct}$ by the average yield for that crop in the district across all years in the data $yield_{dc}$.[4]

[1] The AIS was not conducted in the states of Bihar and Maharastra before 2012. Thus, we exclude these states from our analysis.

[2] The GAEZ dataset reports information on the amount of land – expressed in hectares – farmed with a specific crop in a given cell. The data refers to the baseline year 2000. We focus on the 10 major crops by area harvested in India, namely: rice, wheat, maize, soybean, cotton, groundnut, rape, millet, sugar and sorghum. According to FAOSTAT, the area harvested with these 10 crops amounts to 135.5 million hectares and accounts for 76% of the total area harvested in India in 2000.

[3] As an example, suppose that in district d, 20% of land farmed with rice and 50% of land farmed with wheat are farmed using high-yielding variety seeds. Suppose also that 40% of land in cell i that is part of district d is farmed with rice, while the remaining 60% is farmed with wheat. Under our neutral assignment rule, we assign 38% of land in cell i to high-yielding varieties: $(0.2 \times 0.4) + (0.5 \times 0.6) = 0.38$.

[4] That is, $yield_{dc} = \frac{1}{T} \sum_{t=1}^{T} yield_{dct}$.

The within-district variation generated by our assignment rule is driven by the baseline crop composition of each cell coupled with district-crop level variation in technology adoption and agricultural yield. One potential concern with this assignment rule is that it may generate non-classical measurement error. To see this, let y_i^* be the true level of the share of adoption of a given technology in cell i, and y_i be the imputed measure. We define measurement error in estimated technology adoption by η_i such that $y_i = y_i^* + \eta_i$. After some algebra, it is easy to show that $\eta_i = \sum_{c \in C_i} s_{idc} \times (y_{idc}^* - y_{dc})$, where y_{idc}^* is the true cell-crop share under the technology, y_{dc} is the district-crop level shares obtained from the AIS data, and s_{idc} is the share of cell area farmed under crop c. Letting x_i represent our main treatment variable – i.e. share of non-official language speakers – one would then estimate $\beta = \frac{cov(y_i^*+\eta_i, x_i)}{var(x_i)} = \frac{cov(y_i^*, x_i)}{var(x_i)} + \frac{cov(\eta_i, x_i)}{var(x_i)} = \beta^* + \frac{cov(\eta_i, x_i)}{var(x_i)}$. Any correlation of η with the share of non-state language speakers will bias our estimates of how language barriers affect technology adoption.

To see how various sources of measurement error could affect our estimates, we decompose — without loss of generality — the differences in true cell-crop level shares and observed district-crop shares into a cell-specific, a district-specific, and an idiosyncratic component: $y_{idc}^* - y_{dc} = y_i + y_d + \epsilon_{idc}$. This yields the following expression for the bias in β:

$$cov(\eta_i, x_i) = cov\left(\sum_{c \in C_i} s_{idc}\ y_i, x_i\right) + cov\left(\sum_{c \in C_i} s_{idc}\ y_d, x_i\right) + cov\left(\sum_{c \in C_i} s_{idc}\ \epsilon_{idc}, x_i\right) \quad (3)$$

First, it could be that cells that differ in their share of non-state language speakers are also on a different growth trajectory. This bias is reflected in the first term on the right-hand side of the equation (3). This would happen if, for example, cells with a higher share of non-state language speakers are also cells where farmers grow crops characterized by fast technology adoption. To address this concern, in the paper we show that the share of non-state language speakers is uncorrelated with trends in technology adoption in the five years before the introduction of KCC.

Second, our estimates could be biased downwards if cells with higher non-state speakers also have smaller area farmed under the ten crops considered. This is because under the neutral assignment rule, any changes in district-level shares will be less attributable to cells with lower shares of farmed area. This bias is reflected in the second term on the right-hand side of equation (3). To control for this potential source of bias, our main specification controls for the share of cell's area under the ten crops considered.

In summary, measurement error will have to vary in a very particular way across time, technology and crops to explain our findings. Moreover, the error will have to also vary across spatially adjacent cells that share the same subdistrict borders.

A.A Validation of Technology Adoption Measures and Correlation between Yield Measures

In this section, we validate two of the measures of technology adoption (adoption of HYV seeds and irrigation) using alternative datasets that are publicly available to researchers and that contain information on technology adoption at the village level. In addition, we also investigate the correlation between our main measure of productivity and the three satellite-based proxy measures of productivity discussed in the text.

Information on the use of HYV seeds at village level is seldom available for India. One exception is the ICRISAT Village Dynamics in South Asia (VDSA) dataset, which is based on a household survey that collects information on cultivation practices. The data records the crops farmed by each household, the total area farmed under each crop and how much of the farmed area is cultivated with improved or HYV seed varieties. The finest geographical unit of observation in the VDSA data is a village. The survey covers 17 villages across the five states of Andhra Pradesh, Gujarat, Karnataka, Madhya Pradesh and Maharashtra in 2012 with non-missing information on HYV seeds.[5]

We use information in the VDSA data to calculate the total area farmed in each village under a given crop as well as how much of that area is cultivated using HYV seeds. Similarly, we use the share of area farmed with a given crop in a given cell using the data from the Agricultural Input Survey and the methodology described above. We then map each 10×10 km cell to VDSA villages based on village centroids. This provides us with 30 observations at the cell-crop level for which we observe HYV adoption in both sources. Figure A.1 shows that our measure is positively correlated with the VDSA data at village level: the slope of the line is 1.06 and statistically significant ($t = 4.33$).

We also validate our measure of irrigation using information available in the Village Census of India 2001. The Village Census reports information on area of land irrigated for all Indian villages for the year 2001. We construct a measure of share of irrigated land area for each of our 10×10 km cell by assigning villages to cells based on the geographical coordinates for the centroid of the village. We compare our measure of share of cell area irrigated in the year 2001 against the one reported in the village census data. This provides us with 25,017 observations at the cell level for which we observe share of irrigated land in both the Village Census and with our measure. Figure A.2 shows that our measure is positively correlated with the Village Census measure: the slope of the line is 1.1 and statistically significant ($t = 43.75$).

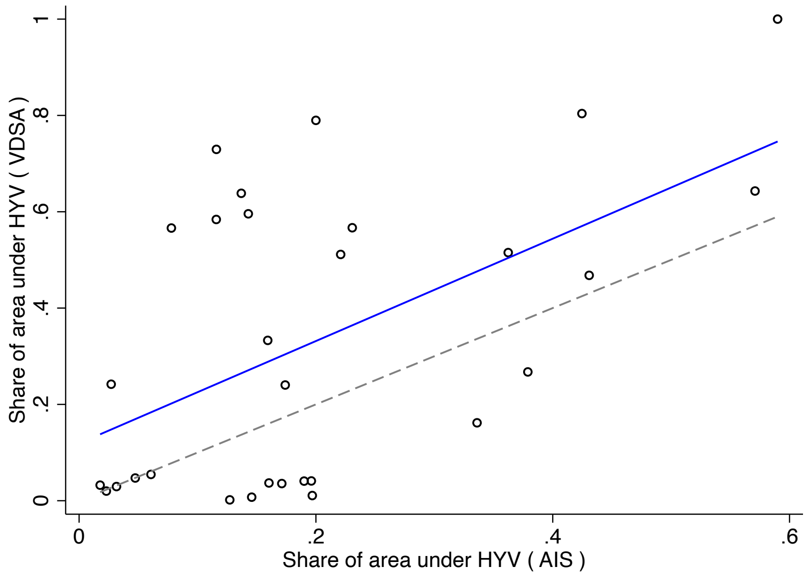
Next, we investigate the correlation between our main measure of productivity - average crop yields from ICRISAT - and the three satellite-based proxy measures of productivity based on the Enhanced Vegetation Index (EVI^{Delta}, EVI^{Max} and $EVI^{Cum.}$).

[5] A potential alternative data source on the use of HYV seeds is the Tamil Nadu Socioeconomic Mobility Survey (TNSMS) conducted by the Economic Growth Center at Yale University. One issue with the TNSMS is that it does not provide village identifiers like VDSA.

Figure A.3 considers the correlation at the district level, while Figure A.4 at the cell level. In all figures we report both the linear fit (conditional on unit of observation and time fixed effects) and the non-parametric visualization of the correlation, obtained by grouping units of observation into equal-sized bins based on their EVI value.

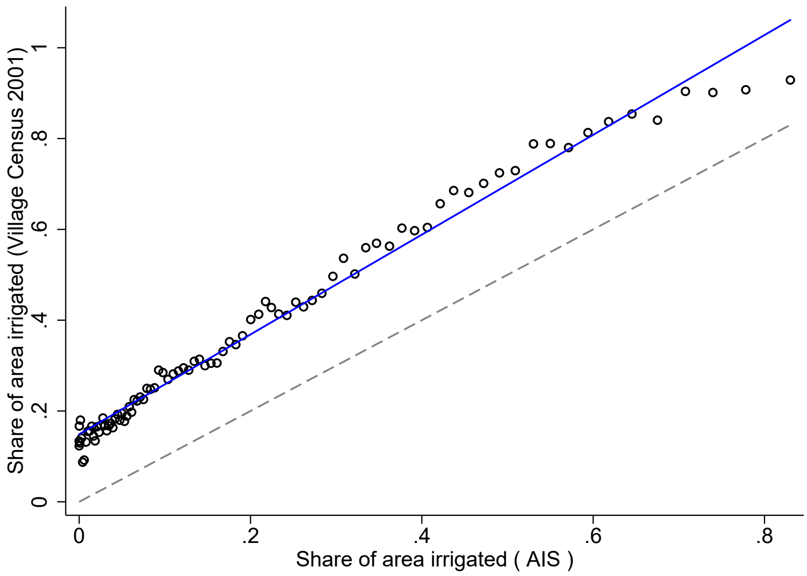
As shown, irrespective of the EVI measure and unit of analysis used, the average crop yields from ICRISAT and satellite-based measure of yields are highly positively correlated, with linear slope coefficients statistically significant at the 99% confidence level in all cases (t-statistic value ranging between 6.71 and 6.88 at the district level, and between 20.70 and 24.03 at the cell level).

Figure A.1: Data Validation: HYV Adoption



Notes: The graph reports the share of crop area under HYV as calculated from ICRISAT VDSA (Village Dynamics in South Asia) micro data against the share of crop area under HYV seeds as calculated from AIS (Agricultural Input Survey). Each dot represents a cell-crop observation for the two measures of share of area under HYV seeds in 2012. The figure has 30 observations and the slope of the line is 1.06 ($t = 4.33$). The dashed gray line is the 45 degree line.

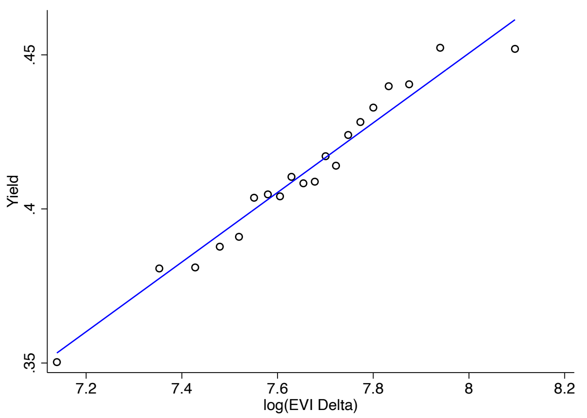
Figure A.2: Data Validation: Share of irrigated area



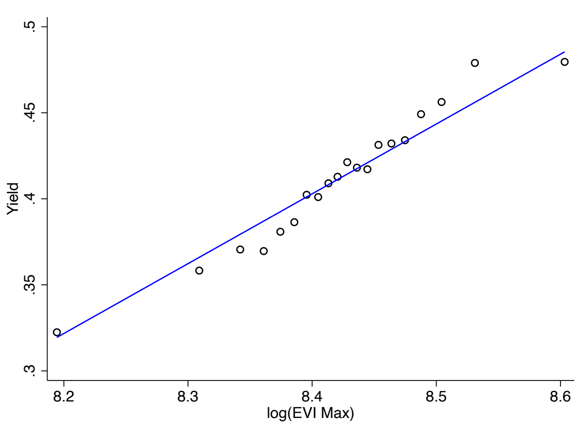
Notes: The graph reports the share of cell area under irrigation as calculated from Villages Census of India 2001 against the share of cell area under irrigation as calculated from AIS (Agricultural Input Survey) 2001. Each dot has 1% of observation based on the share of irrigated area measured through AIS and represents the average of the two measures of share of area under irrigation in 2001. The slope of the line is 1.1 ($t = 43.75$). The dashed gray line is the 45 degree line.

Figure A.3: Correlation between different productivity measurements: District level

(a) Correlation between Yield and $log(EVI^{Delta})$



(b) Correlation between Yield and $log(EVI^{Max})$



(c) Correlation between Yield and $log(EVI^{Cum.})$

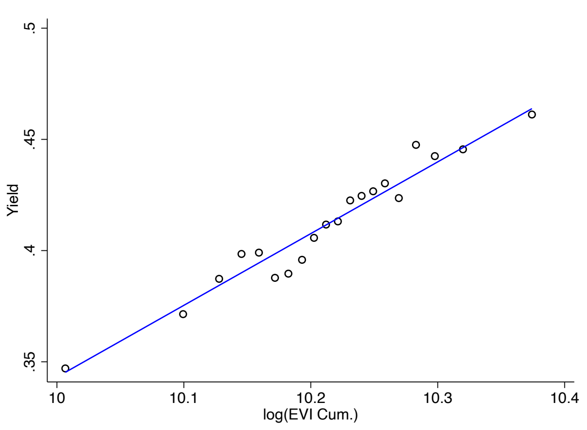
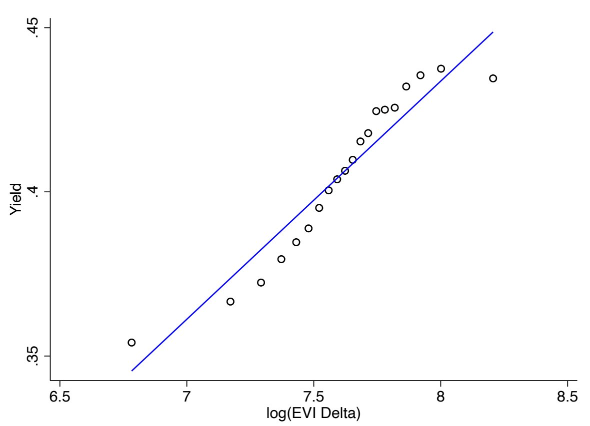
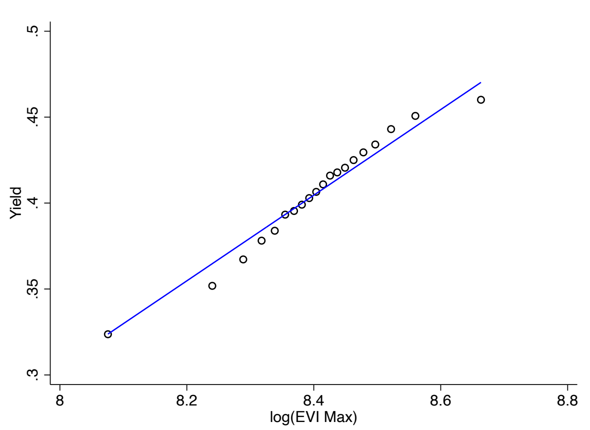


Figure A.4: Correlation between different productivity measurements: Cell level

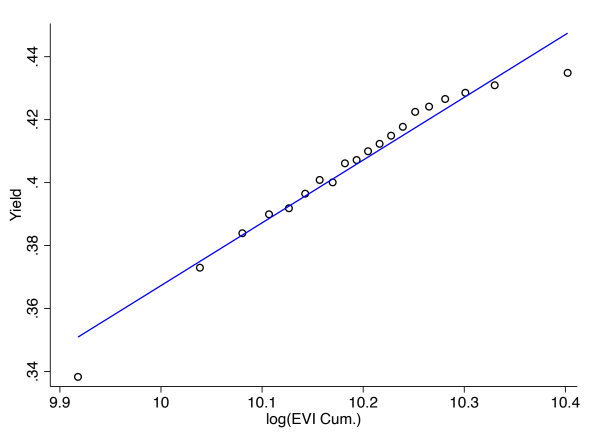
(a) Correlation between Yield and $log(EVI^{Delta})$



(b) Correlation between Yield and $log(EVI^{Max})$



(c) Correlation between Yield and $log(EVI^{Cum.})$



B Classification of KCC calls

In this section we provide examples of calls on agricultural technologies made by farmers to the Kisan Call Centres (KCC). We classify as calls about agricultural technologies those in which farmers ask questions about seeds, fertilizers, pesticides and irrigation. We extract this information from farmers' queries ("QueryText") and agronomists' answers ("Answer").

Panel A of Table B.1 refers to calls about seeds. These include (i) calls asking directly about hybrid varieties related to a crop and (ii) queries or answers about specific high-yielding seed varieties. The questions are crop-, period- and area-specific. In the examples shown, farmers call from Haryana and Punjab, two of the country's major wheat and rice producing states, respectively, to ask about high-yielding seed varieties at the beginning of their respective growing seasons, October and June.

Panel B refers to calls about fertilizers. We classify as calls on fertilizers: (i) calls seeking general information on fertilizer dosage; (ii) calls directly asking remedies for nutrient deficiencies in crops; (iii) queries or replies based on required dosage of specific fertilizers, *e.g.* N-P-K or Urea; (iv) calls seeking information on plant growth regulators, seed treatment or solution to leaf drop. In many calls farmers ask about the dosage of specific fertilizers, as reported in the examples below.

Panel C covers calls about pesticides. We classify as calls on pesticides: (i) calls seeking specific information on pesticides; (ii) agronomists suggesting the use of certain pesticides like Quinalphos and Chlorpyriphos[6]; (ii) calls asking for solutions to pest infections; (iii) calls related to plant protection; (iv) inquiries about weed control. In the examples below, farmers inquire about how to response to specific pests, from leaf-folders to termites.

Finally, Panel D refers to calls about irrigation and water management. To classify calls on irrigation, we use questions from farmers seeking information on: (i) irrigation practices; (ii) water management in the field. Most of the calls concern the suitable time for particular stages of irrigation, as shown in the examples.

[6] Quinalphos is a pesticide widely used in India for wheat, rice, coffee, sugarcane, and cotton. Chlorpyriphos is a pesticide used to kill a number of pests, including insects and worms.

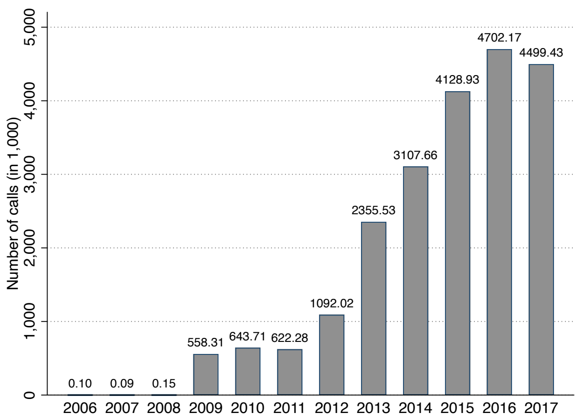
Table B.1: Examples of Calls on Agricultural Technologies

Date	State	subdistrict	QueryText	Answer
			Panel A: Calls on seeds	
2012-10-05	Haryana	Naraingarh	Information on improved varieties of wheat	w.h.-1105,w.h.d.-948, w.h-1025,w.h.-416,c.-316
2011-06-18	Punjab	Dasuya	Information on improved varieties of basmati rice	Basmati-386, Pusa Basmati No-1, Basmati-370
			Panel B: Calls on fertilizers	
2012-01-16	Chattisgarh	Manpur	To know about fertilizer in wheat at tillering	Apply 30kg.urea/acre at tillering stage
2012-02-10	Tamil Nadu	Kuttalam	Top dressing fertilizer management for rice	Apply 25 kg Urea + 15 kg Potash and 5 kg Neemcake
			Panel C: Calls on pesticides	
2012-03-14	Tamil Nadu	Tiruvallur	How to control rice leaf-folders	Spray Quinalphos at 2ml/lit
2012-02-10	Uttar Pradesh	Derapur	Termite in sugarcane	Apply Chlorpyriphos at 4lit/hac with irrigation water
			Panel D: Calls on irrigation	
2011-05-13	Haryana	Jagadhri	Time of first irrigation in cotton?	After 45 days of sowing time
2011-01-08	Rajasthan	Anupgarh	Tell me interval of time of irrigation in mustard	40-45 days

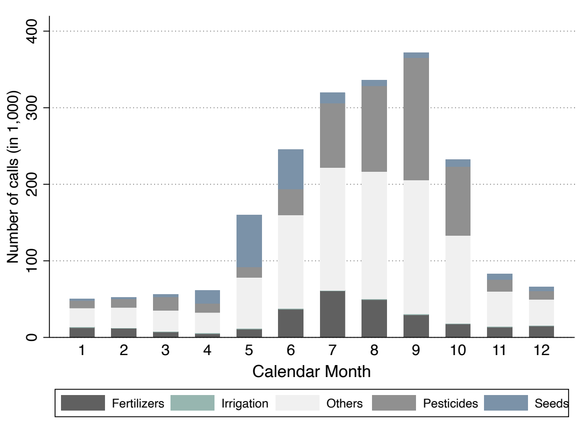
C Appendix Figures and Tables

Figure C.1: Calls to Kisan Call Centers: 2006-2017

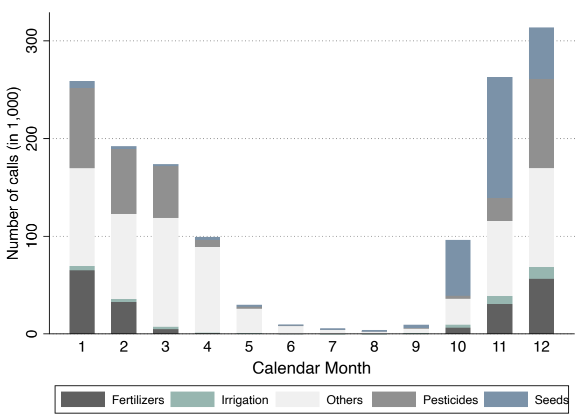
(a) Total number of calls



(b) Calls about rice (*kharif* season) by calendar month and topic

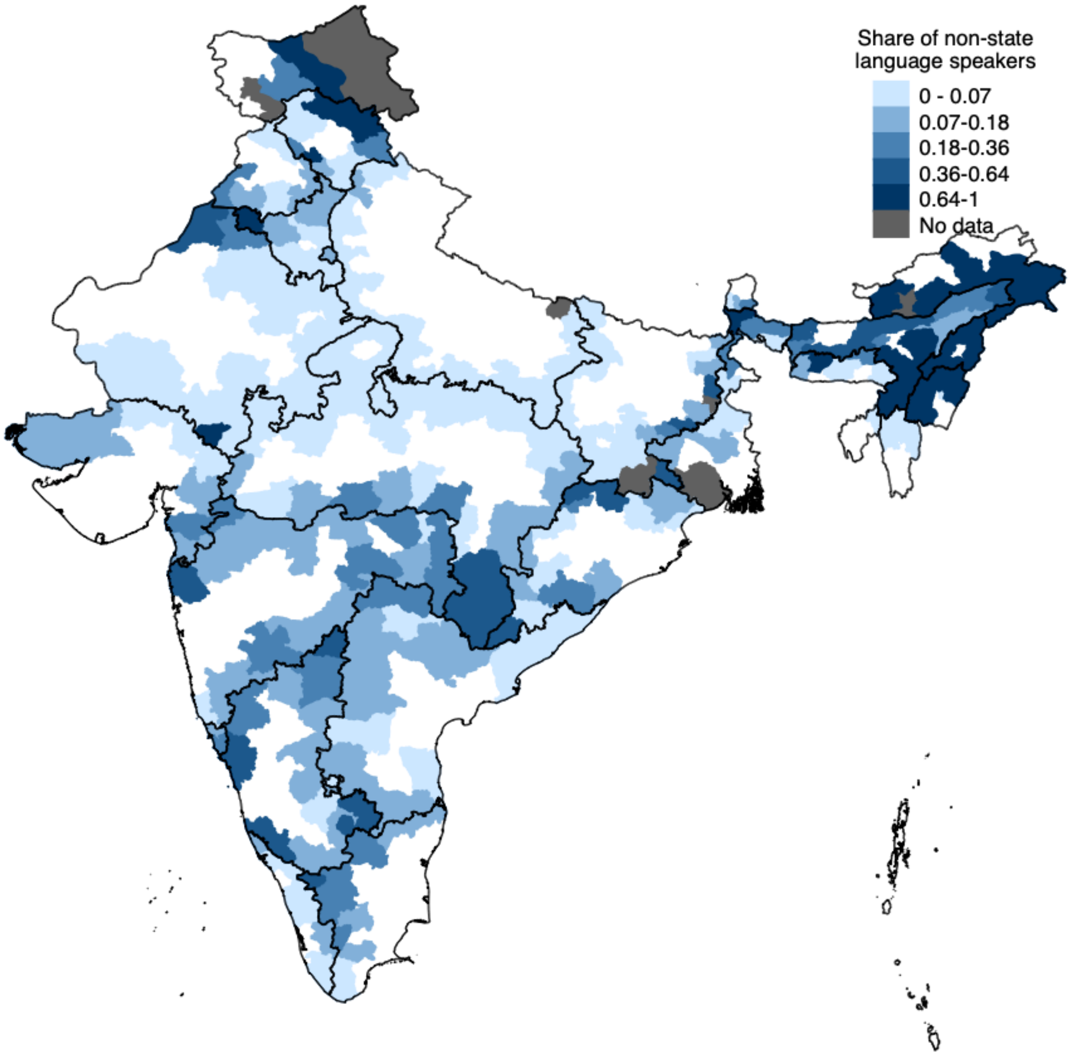


(c) Calls about wheat (*rabi* season) by calendar month and topic



Notes: Source: Kisan Call Center, Ministry of Agriculture

Figure C.2: District variation in
Non-state official language speakers



Notes: Source: 2011 Population Census of India. Speakers identified by their primary language. The figure report the share of non-state language speakers out of the official language speakers across Indian districts that are geographically contiguous to borders shared by two states.

Table C.1: Robustness: Alternate Productivity Measures

Outcome:	Yield	$\log(EVI^{Delta})$	$\log(EVI^{Max})$	$\log(EVI^{Cum.})$
	(1)	(2)	(3)	(4)
Non-state official language speakers (%) × Post	-0.010***	-0.020	-0.011*	-0.008
	[0.003]	[0.022]	[0.006]	[0.006]
Observations	129,843	129,843	129,843	129,826
R-squared	0.990	0.909	0.931	0.962
Mean	0.383	7.462	8.358	10.185
Cell f.e.	✓	✓	✓	✓
Subdistrict border × Time f.e.	✓	✓	✓	✓
District × Time f.e.	✓	✓	✓	✓
Majority language × Time f.e.	✓	✓	✓	✓

Notes: The table reports the estimates across various measures of agricultural productivity using specification 1. Column (1) produces the estimates from Table III that using our main measure of agricultural productivity. Columns (2)-(4) report estimates using alternative measures of productivity from the Enhanced Vegetation Index (EVI), an index of intensity of vegetation cover estimated by the US Geological Survey using the Moderate Resolution Imaging Spectro-radiometer (MODIS) aboard NASA's Earth Observing System-Terra satellite. Vegetation indices such as EVI exploit plant reflectance of electromagnetic radiations to quantify vegetation greenness in an area, whose spatial distribution is estimated from satellite images. Standard errors accounting for spatial correlation within a $10km$ distance from each cell using the correction proposed in Conley (1999) are reported in brackets. Significance levels: *** p<0.01, ** p<0.05, * p<0.1 .

Table C.2: Robustness: Alternative distances to border

Outcome:	All Calls per 100 farmers (1)	HYV Adoption (2)	Yield (3)
Distance to border: 30 km			
Non-state official language speakers (%) × Post	-3.019	-0.021**	-0.007**
	[2.035]	[0.008]	[0.003]
Observations	76,932	24,409	89,606
Distance to border: 35 km			
Non-state official language speakers (%) × Post	-3.460**	-0.018***	-0.005*
	[1.719]	[0.007]	[0.003]
Observations	86,916	27,576	101,471
Distance to border: 40 km			
Non-state official language speakers (%) × Post	-4.560***	-0.018***	-0.006**
	[1.663]	[0.006]	[0.003]
Observations	96,576	30,648	112,650
Distance to border: 45 km			
Non-state official language speakers (%) × Post	-4.237***	-0.020***	-0.009***
	[1.446]	[0.006]	[0.003]
Observations	105,564	33,504	123,302
Distance to border: 50 km			
Non-state official language speakers (%) × Post	-3.870***	-0.021***	-0.010***
	[1.324]	[0.006]	[0.003]
Observations	113,868	36,139	129,843
Distance to border: 55 km			
Non-state official language speakers (%) × Post	-3.319***	-0.020***	-0.008***
	[1.193]	[0.005]	[0.003]
Observations	122,100	38,773	142,702
Distance to border: 60 km			
Non-state official language speakers (%) × Post	-2.859**	-0.021***	-0.010***
	[1.154]	[0.005]	[0.003]
Observations	129,192	41,028	150,823
Cell f.e.	✓	✓	✓
Subdistrict border × Time f.e.	✓	✓	✓
District × Time f.e.	✓	✓	✓
Majority language × Time f.e.	✓	✓	✓

Notes: The table reports the robustness of the main estimates from using specification 1 with alternative distances to state borders ranging from 30*km* to 60*km*. Standard errors accounting for spatial correlation within a 10*km* distance from each cell using the correction proposed in Conley (1999) are reported in brackets. Significance levels: *** p<0.01, ** p<0.05, * p<0.1 .

Table C.3: Robustness: Alternative Clustering
Conley Standard Errors

Outcome:	All Calls per 100 farmers	HYV Adoption	Yield
	(1)	(2)	(3)
Non-state official language speakers (%) × Post	-3.870	-0.021	-0.010
Spatial Correlation, threshold: 10 km (Baseline)	[1.324]***	[0.006]***	[0.003]***
Spatial Correlation, threshold: 25 km	[1.352]***	[0.006]***	[0.003]***
Spatial Correlation, threshold: 50 km	[1.406]***	[0.007]***	[0.003]***
Spatial Correlation, threshold: 100 km	[1.573]**	[0.007]***	[0.004]***
Spatial Correlation, threshold: 150 km	[1.622]**	[0.007]***	[0.004]**

Notes: Significance levels: *** p<0.01, ** p<0.05, * p<0.1 .

Table C.4: Robustness to Controlling for Religion Groups and Majority Language

Outcome:	All Calls per 100 farmers	HYV Adoption	Yield
	(1)	(2)	(3)
Non-state official language speakers (%) × Post	-2.246**	-0.024***	-0.009**
	[1.086]	[0.007]	[0.004]
Observations	92,544	29,070	102,099
R-squared	0.563	0.993	0.987
Mean	2.025	0.284	0.383
Cell f.e.	✓	✓	✓
Subdistrict border × Time f.e.	✓	✓	✓
District × Time f.e.	✓	✓	✓
Majority language × Time f.e.	✓	✓	✓
Majority religion × Time f.e.	✓	✓	✓

Notes: The table reports the robustness of the main regression specification to the inclusion of majority religion fixed effects. Majority religion is defined as the religious group that forms the majority in a cell and is obtained from the 2011 Census data. There are 6 main religious groups across cells in our sample: Buddhists, Christians, Hindus, Muslim, Sikhs, Others. Standard errors accounting for spatial correlation within a $10km$ distance from each cell using the correction proposed in Conley (1999) are reported in brackets. Significance levels: *** p<0.01, ** p<0.05, * p<0.1 .

Table C.5: Robustness: Alternative Geographical Units of Observation

Outcome:	All Calls per 100 farmers	HYV Adoption	Yield
	(1)	(2)	(3)
Panel A: District-Crop as unit of analysis			
Non-state official language speakers (%) × Post		-0.133**	-0.131**
		[0.063]	[0.051]
Observations		4,727	24,928
R-squared		0.762	0.321
District × Crop f.e.		✓	✓
Border × Time f.e.		✓	✓
Crop × Time f.e.		✓	✓
Census Controls × Time f.e.		✓	✓
Panel B: Subdistrict as unit of analysis			
Non-state official language speakers (%) × Post	-0.809**	-0.013*	-0.019**
	[0.393]	[0.007]	[0.007]
Observations	17,556	4,798	17,420
R-squared	0.886	0.998	0.996
Subdistrict f.e.	✓	✓	✓
Crop Share × Time f.e.	✓	✓	✓
Subdistirct border × Time f.e.	✓	✓	✓
District × Time f.e.	✓	✓	✓
Census controls × Time f.e.	✓	✓	✓

Notes: The table conducts the analysis at alternative geographical units. Panel A conducts the analysis at the district-crop level. The effect on calls cannot be estimated with this specification because a large share of call-level data does not report information on the crop the farmer asks information about. All columns include district-crop fixed effects, common state border-time fixed effects, state-time fixed effects, crop-time fixed effects, and district-level controls of share of scheduled caste population, literacy rate, total population, number of agricultural societies, and number of agricultural workers, all interacted with time fixed effects. Standard errors clustered at the district level are reported in brackets. Panel B conducts the analysis at the subdistrict level. All columns include subdistrict fixed effects, subdistrict border-time fixed effects, district-time fixed effects, share of subdistrict area farmed with 10 crops at baseline interacted with time-fixed effects, and share of subdistrict population in agriculture from the 2001 Census interacted with time-fixed effects. Standard errors clustered at subdistrict level are reported in brackets. Significance levels: *** p<0.01, ** p<0.05, * p<0.1 .