May 2017 No.331

Linguistic Distance and Market Integration in India

James Fenske and Namrata Kala

WORKING PAPER SERIES

Centre for Competitive Advantage in the Global Economy

Department of Economics

LINGUISTIC DISTANCE AND MARKET INTEGRATION IN INDIA

JAMES FENSKE† AND NAMRATA KALA⋆

ABSTRACT. We collect data on grain and salt prices, as well as language, for more than 200 South Asian markets in the 19th and early 20th centuries. Conditional on a rich set of controls and fixed effects, we find that linguistically distant markets are less integrated as measured by the degree of price correlation. While linguistically distant markets exhibit greater genetic distance, greater differences in literacy, and fewer railway connections, these factors are not sufficient statistics for the negative correlation between linguistic distance and market integration. Our results indicate that a one standard deviation increase in linguistic distance predicts a reduction in the price correlation between two markets of 0.121 standard deviations for wheat, 0.167 standard deviations for salt, and 0.088 standard deviations for rice. These differences are substantial relative to other factors such as physical distance that hinder market integration.

1. INTRODUCTION

Well-functioning markets are important drivers of allocative efficiency and welfare, and markers of economic development (Shiue and Keller, 2007; Studer, 2008). Although such language barriers have been stressed in the macroeconomic literature as inhibiting trade and the diffusion of technology (Guiso et al., 2009; Spolaore and Wacziarg, 2009), the role of these variables in market integration within countries or within economies, particularly in the developing world, has received comparatively little attention. This is despite the sizable economic impacts that these barriers can have in other contexts (Ashraf and Galor, 2013; Spolaore and Wacziarg, 2016). In this paper, we consider the economy of colonial India, for which consistent price data are available across many years, and in which the a large number of dissimilar languages prevail. In particular, we ask: do market pairs that are more linguistically distant display less market integration, conditional on physical distance and other measures of dissimilarity?

†UNIVERSITY OF WARWICK
⋆HARVARD UNIVERSITY
E-mail addresses: J.Fenske@warwick.ac.uk, kala@fas.harvard.edu.
Date: May 10, 2017.
We are grateful to Martin Fiszbein, Marc Klemp, and to audiences at the Association for the Study of Religion, Economics, and Culture, George Mason University, the University of Manchester, and the University of Warwick for their comments. Extra thanks are due to Marlous van Waijenburg for sharing additional price data with us, and to Paradigm Data Services (inquire@pdspl.com), Connie Yu and Mina Rhee for their assistance in data entry.

In this paper, we collect data on grain and salt prices for 206 South Asian markets in the 19th and early 20th centuries from *Wages and Prices in India*. These markets span the territories of modern-day Bangladesh, Burma, India, and Pakistan. We merge these markets to populations by language collected from the 1901 colonial census of India. We next join these data to language trees from *Ethnologue*. These language trees allow us to compute cladistic distances between the languages in the data and, as a result, between every market. Taking the correlation coefficient between the price series at a pair of markets i and j, we show that, conditional on physical distance, religious distance, dissimilarities in geography, and fixed effects for markets i and j, prices at i and j are less correlated if i and j are more linguistically distant. Our estimates suggest that two markets with unrelated languages will, compared to two markets sharing a common tongue, have correlation coefficients that are 0.067 less in the case of wheat, 0.224 less in the case of salt, and 0.035 less in the case of rice, relative to means of 0.81 (wheat), 0.33 (salt) and 0.81 (rice) across all market pairs in the data.

In assessing the mechanisms that link linguistic distance to market integration, we turn to both the economic literature and to the history of colonial India. Recent work in economics has stressed several mechanisms that could help explain our results. Language barriers may proxy for more general barriers to the transmission of vertical traits (Spolaore and Wacziarg, 2009, 2016); they may capture differences in tastes, and hence the presence of markets (Atkin, 2013, 2016); they may affect the costs of information transmission and coordination (Gomes, 2014); they may otherwise affect trade costs through interaction, migration, business connections, conflict, or xenophobia (Bai and Kung, 2014; Falck et al., 2012; Lameli et al., 2015; Laval et al., 2016; Rauch and Trindade, 2002); they may work through costs of language acquisition and education acquisition (Isphording and Otten, 2014; Jain, 2015; Laitin and Ramachandran, 2016; Shastry, 2012); they may correlate with common preferences for public goods, redistribution, and infrastructure (Desmet et al., 2016a, 2012, 2015). In the secondary historical data, migrant communities that share a common language (Markovits, 2008) and linguistic barriers to migration (Collins, 1999) feature prominently.

In order to assess which of these explanations may account for our results, we assemble data from a wide range of primary and secondary sources. We show that market pairs that are more linguistically distant from each other are also more genetically distant, but that this summary measure of barriers to the diffusion of technological and institutional innovations is not itself a sufficient statistic for the coefficient on linguistic distance. We find little evidence that linguistic distance predicts missing markets or fewer shared trading communities. Historic differences in literacy across market pairs do correlate with linguistic distance, but do not fully account for its correlation on price

integration. More linguistically-similar markets spent more years both connected to the colonial railway system, but this too does not explain away the effect. So: while linguistic distance may have operated in part as a marker of other population differences, as a barrier to the acquisition of similar levels of human capital, and as a barrier to the co-acquisition of public goods that facilitated trade, no one of these mechanisms can fully account for the barriers given by of linguistic cleavages.

We demonstrate robustness to selection on unobservables, show that our results hold across several other crops, restrict our sample to modern India, use alternative measures of linguistic distance and alternative functional forms, and remove outliers, among other exercises.

1.1. **Contribution.** Our paper contributes principally to two literatures. The first investigates the role of linguistic distance in particular, and cultural distances more broadly, in shaping economic outcomes. Linguistic, cultural, and ethnic cleavages are strong predictors of civil conflict (Desmet et al., 2012; Esteban et al., 2012), trade (Hutchinson, 2005), health (Gomes, 2014), public goods provision (Dickens, 2016), and redistribution (Desmet et al., 2016b) in modern data. More generally, linguistic, religious, and cultural distances across societies correlate with ancestral distance and predict a wide range of economic outcomes (Spolaore and Wacziarg, 2016).

Second, we contribute to a literature on market integration and trade. This is important, because market fragmentation can increase volatility, and greater price volatility impedes development (Jacks et al., 2011). Following on works such as Persson (1999) and Shiue and Keller (2007), several contributions in economic history have measured price integration across markets in order to compare levels of economic development across regions (Federico, 2011; O'Rourke and Williamson, 2002; Studer, 2008). Other studies have used historical price series to measure the responsiveness of prices and welfare measures to variables such as weather shocks and transportation infrastructure (Andrabi and Kuehlwein, 2010; Jia, 2014; Waldinger, 2014). More generally, our work is related to a broader literature on the evolution of trade and market integration throughout history (Estevadeordal et al., 2003; Jacks et al., 2008; Pascali, 2016).

We make several contributions to these literatures. We add cultural costs like linguistic distance to the determinants of market integration, and explore mechanisms via which these affect market functioning. In a literature in which intra-country evidence on linguistic barriers to trade integration is itself rare, we provide one of the first studies to investigate this question in a developing country context. We make a substantial data contribution, digitizing detailed language data from the colonial census and price data

covering a wider set of markets and commodities (68,181 observations) than those used by Allen (2007), Andrabi and Kuehlwein (2010) or Studer (2008).[1]

The most similar studies to ours, then, are Falck et al. (2012) and Lameli et al. (2015). These papers use dialect similarity within Germany to predict intra-regional trade and migration flows. One difference of our work and theirs is that these papers do not include genetic distance in their estimating equations. Since linguistic may proxy for a broader set of inter-population differences, it is important to assess the degree to which basic summary markers such as genetic distance may account for the coefficient on linguistic distance.[2] We focus on differences across languages, rather than dialects. Further, we test whether this correlation is a result of or mitigated by transport investment choices. Finally, we provide evidence from a large and multilingual developing country, covers a longer time period, examines price integration as an outcome, and uses a more spatially disaggregated unit of analysis.

The rest of the paper proceeds as follows. Section 2 outlines the identification strategy and data sources. Section 3 presents the main results. Section 4 discusses additional results and empirically assesses mechanisms. Section 5 outlines robustness checks. Section 6 concludes.

2. Identification strategy and Data

2.1. **Identification strategy.** In this paper, we use price data covering M South Asian markets. Each observation is a market-pair, indexed ij. For product p, traded between markets i and j, we estimate:

$$\rho_{ij}^{p} = \beta^{p} LinguisticDistance_{ij} + x_{ij}^{p\prime}\gamma^{p} + \delta_{i}^{p} + \eta_{j}^{p} + \epsilon_{ij}^{p}. \tag{1}$$

In (1), ρ_{ij}^{p} is the correlation coefficient for the price of p between markets i and j. $LinguisticDistance_{ij}$ is described below, and captures linguistic distance between the two markets. x_{ij}^{p} is a vector of controls. We use this to account for a wide set of dissimilarities between i and j that may correlate with linguistic distance and with the degree of price integration. In our baseline estimations, x_{ij}^{p} includes a constant, minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar,

[1] In particular, we have chosen to make our raw data on prices, languages, and the spread of the railway network available online: see `www.jamesfenske.com`.

[2] See Giuliano et al. (2014) as an example for trade between countries.

tea, wetland rice, white potato, wheat, and tomato. δ_i^p and η_j^p are fixed effects for market i and market j. The sample is all market pairs ij such that $i \neq j$, $i > j$ and there are sufficient observations to compute ρ_{ij}^p. That is, we have at most $\frac{M^2-M}{2}$ observations in any one regression. We cluster standard errors by both market i and market j in the baseline (Cameron et al., 2011).

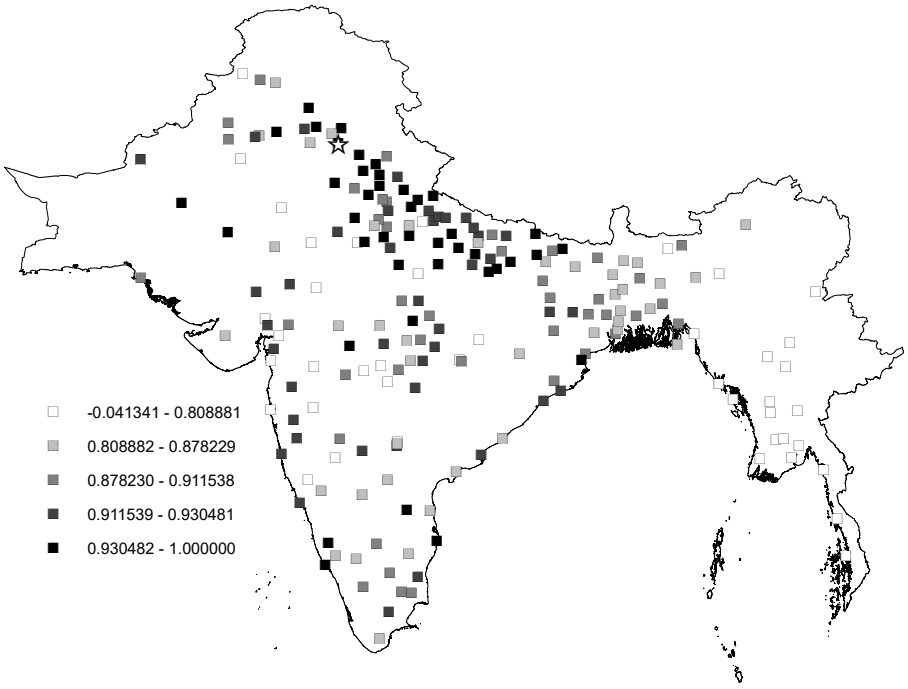
2.2. **Data.** We use several sources of data. Below, we discuss our sources for prices in colonial India, for linguistic distance across markets, and for our additional controls.

2.2.1. *Prices.* Our data on prices are taken from three editions (1921, 1907, and 1885) of *Wages and Prices in India*. These are initially in reported in sers (~ 1.15 kg) per rupee: we invert this measure to obtain nominal prices. For 206 markets in modern-day Pakistan, India, Bangladesh, and Burma, these data provide prices for several crops: Arhar Dal, Barja, Barley, Gram, Jawar, Kangni, Maize, Marua, Rice, Salt, Wheat, Bulrush Millet and similar, Great Millet and Similar, and Lesser Millets. In most of our results, we focus on the three most commonly-reported prices: rice, wheat, and salt. We do, however, show that estimates of (1) with several other crops produce similar estimates. The price data cover the period 1861 through 1921, with many markets entering our data for the first time in 1869. While the data collection methods differed across markets in early years, from 1872 onwards these are based on uniform fortnightly returns of retail prices. When a price is reported more than twice at markets i and j in the same year, we can compute a correlation coefficient for that product for the ij pair. This quantity, ρ_{ij}^p, is our principal dependent variable.

In Figure 1, we provide intuition for our results by mapping the correlation between the price of rice in a single market, the largely Punjabi-speaking city of Ludhiana, with the price of rice in all other markets in our data. It is clear from the figure that rice prices track those in Ludhiana more closely in regions that speak more closely-related languages such as Hindi and Gujarati and less closely in regions that speak more distantly-related languages such as Burmese and Telugu. These regions are, however, also closer in physical proximity to Ludhiana, and many of the markets that most closely track prices in Ludhiana lie on the Indo-Gangetic Plain. So: our analysis relies on estimation of (1) in order to demonstrate that the correlation between linguistic distance and price integration cannot be explained away by other observable differences in proximity or geography.

The secondary literature on Indian history provides some information on how these local prices of foodgrains were determined. Production varied by region, and grains were largely consumed domestically. For example, 70% of wheat acreage in 1919 was in the Punjab and United Provinces, of which only 5% was exported beyond India in 1895. By contrast, 70% of rice acreage in 1919 was in Bengal, Bihar, Orissa, and Madras, and

FIGURE 1. Ludhiana: Rice price correlations



only 7% was exported in 1895 (Andrabi and Kuehlwein, 2010). The non-monetary sector of the economy was large (Kumar, 1983), even in 1950 (Chandavarkar, 1983).

At the start of our period, trade costs were high. Overland transport was expensive, along dilapidated roads and with food grains carried by oxen in carts or on back loads (Bhattacharya, 1983). In Western India, trade was largely carried out by donkey, camel and bullock, and there were few constructed roads (Divekar, 1983). Intraregional trade in low-value commodities was possible along rivers but access to this trade was spatially limited (Derbyshire, 1987). Trade by bullocks could, in a year, cover the distance that a railway would later cover in a week (McAlpin, 1974). Caravans carried cotton and grain where a lack of roads made wheeled transportation difficult (Roy, 2012). Large-scale, long-distance shipments of grain were generally unprofitable (Hurd, 1975). The costs of overland transport limited market integration (Kessinger, 1983). Integration of the transport network was still not sufficient to alleviate food shortages during the 1870s (McAlpin, 1974).

River transport of food grains along the Ganges and its tributaries favored larger traders (Bhattacharya, 1983). Commercial activity was greater where rainfall was more abundant and reliable (Derbyshire, 1987). Migration rates were low and wage convergence over the nineteenth century was slow (Collins, 1999). The commercial orbit of the United

Provinces was constrained in its geographical scope by speed, cost, and seasonality (Derbyshire, 1987).

These costs fell during the time period we cover. The telegraph network spread through India in the 1850s and 1870s (Collins, 1999). Increasing commercialization was aided by the replacement of the fragile military occupation with settled governance, a growing market for raw materials in Europe, and infrastructural improvements such as canal irrigation, metalled roads, and railway construction (Derbyshire, 1987; Kumar, 1983). The railways in particular reduced price dispersion across markets (Hurd, 1975). Price dispersion fell more rapidly for cash crops such as cotton than for food grains (McAlpin, 1974). Andrabi and Kuehlwein (2010) find evidence of trade in grain from districts without railroads to neighboring districts that were connected to this network.

Bhattacharya (1983) describes local market places in Eastern India. Farmers might sell directly to consumers and middlemen in small quantities. Itinerant traders made small profits exploiting price differences within limited areas. Large farmers might serve as links between village markets and larger towns by buying grain from smaller farmers through credit contracts, holding stock while waiting for a favorable market, and taking grain to the mart or river mart offering the best price. Merchants' agents played a similar role. In larger towns, trade was stratified into retail sellers, wholesale merchants, and those who bought from wholesalers and sold to retailers. Divekar (1983), Kumar (1983), and Kessinger (1983) provide similar descriptions for other regions of India in the first half of the nineteenth century.

Later in the century commission agents and buyers' agents operated in towns that contained railway stations and banks (Roy, 2014). They owned capital such as carts, grain pits and warehouses. Commission agency and auction-type sales were prevalent, and there was little evidence of forward trade. Company agents contracted with farmers in the villages, while landlords and others lent money to these farmers and were repaid in grain they also sold to the commission and buyers' agents. In more remote areas, itinerant traders, including peasants, brought crops to bazaars. Europeans were largely absent from this trade.

Generally, prices in local markets correlated with fluctuations in the overall Indian money supply (Adams and West, 1979). Prices were typically lower in producing regions (Andrabi and Kuehlwein, 2010). On average, prices rose slowly through the nineteenth century and rapidly during the First World War (McAlpin, 1983).

2.2.2. *Linguistic distance.* To compute linguistic distances between the markets in our data, we use two additional data sources. These are the 1901 Census of India and version 19 of the *Ethnologue* Global Dataset. For each district that existed in 1901, the census data report the number of speakers of each language. For example, the three most

commonly-spoken languages reported for Ludhiana District are "Punjabi" (665,476), "Hindostani" (2,970), and "Kashmiri" (1,224). We assign each market in the data to the language composition of the district that contained it in 1901. For consistency with the *Ethnologue* data on distances, we aggregate these to the level of ISO language codes. For Ludhiana, the three most commonly-spoken languages become pan, hin, and kas.

To compute the distances between these languages, we turn to *Ethnologue*. Every language in this source is categorized using a language tree. For example, Punjabi is coded as Indo-European, Indo-Iranian, Indo-Aryan, Intermediate Divisions, Western, Panjabi. The maximum number of branches is 15. These classifications are based on several sources. The most important is Frawley (2003). Following Esteban et al. (2012), we take the distance d_{ij} between any two languages i and j as:

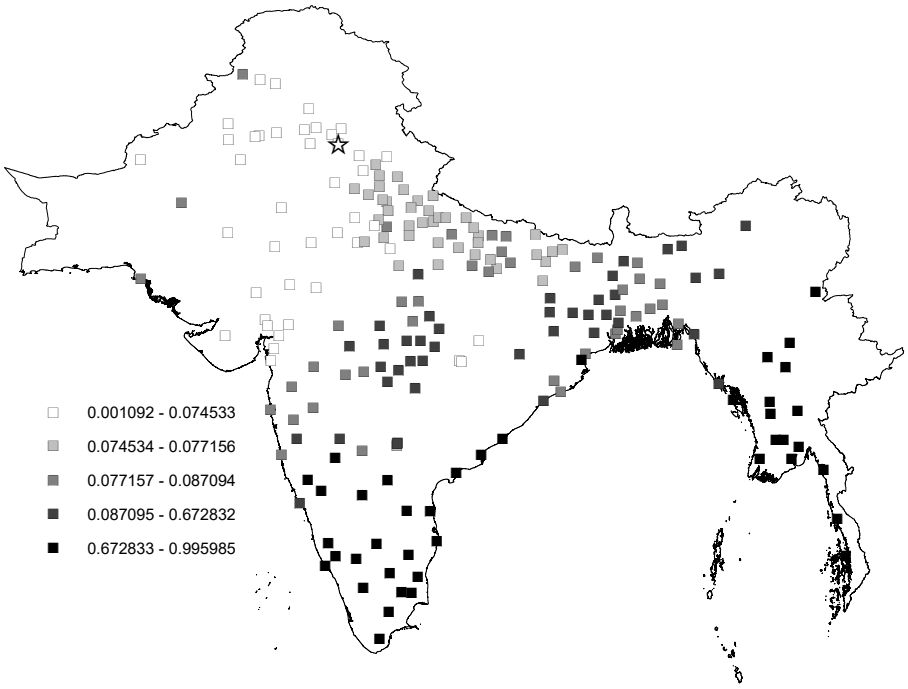
$$d_{ij} = 1 - \left(\frac{SharedBranches}{15}\right)^{\delta}. \tag{2}$$

Similarly following Esteban et al. (2012), we choose $\delta = 0.05$ as a baseline and use $\delta = 0.5$ for robustness. "Cladistic" measures such as this have become widely used in economics (Desmet et al., 2012; Gomes, 2014). Although alternative distance measures exist based on phonetic similarity of languages (Dickens, 2016), these are not feasible given the large number of languages in our data. To aggregate these to distances between markets, abusing notation, and given population shares of languages i and j in each district 1 and 2 of s_{1i} and s_{2j}, we follow Spolaore and Wacziarg (2009) and compute linguistic distance between districts as:

$$LD_{12} = \sum_i \sum_j (s_{1i} \times s_{2j} \times d_{ij}). \tag{3}$$

In Figure 2, we map the linguistic distances between every district in our data and Ludhiana. While it is evident that the markets at which languages more closely related to Punjabi are spoken are geographically close to Ludhiana, it is also clear that this correlation of linguistic and physical distance is not perfect. Distances change relatively rapidly over space when the linguistic composition of the population similarly changes rapidly. Further, regions that are relatively similar in physical distance can be quite dissimilar in their linguistic distance. Punjabi and Bengali, for example, both share the branches Indo-European, Indo-Iranian, and Indo-Aryan. Punjabi and Tamil, by contrast, share no branches, as Tamil is a Dravidian language. And yet the distance between the Punjab and Bangladesh is not markedly different than the distance between the Punjab and Tamil Nadu. The log distance in km between Ludhiana and Dacca is 7.40, whereas it is 7.76 between Ludhiana and Madurai.

FIGURE 2. Ludhiana: Linguistic distances



2.2.3. *Additional controls.* Some of our control variables are computed directly. Distance in km is computed using the latitude and longitude of the market. Both coastal and both connected by the same river are computed in ArcMAP using a shapefile of district boundaries. Minimum year, maximum year, and number of common observations are computed directly from the price data.

The same province indicator codes markets to the provinces that contained them in 1901. The religious distance indicator is computed using the same equation as (3), taking the religious composition of each district as reported in Table 8 of the 1921 Census (Literacy By Religion). We assume that the distance d_{ij} between any religion i and j is 1 if $i \neq j$ and 0 if $i = j$.

Data on land quality is taken from Ramankutty et al. (2002) and has been used in several economic studies, such as Michalopoulos (2012) and Ashraf and Galor (2011).[3] It is an index based on soil and climate characteristics and is not particular to any one type of agriculture. Ruggedness is the measure of terrain ruggedness initially introduced by Nunn and Puga (2012).[4] Our measure of malaria prevalence was originally created

[3] https://nelson.wisc.edu/sage/data-and-models/atlas/maps.php?datasetid=19&includerelatedlinks=1&dataset=19

[4] http://diegopuga.org/data/rugged/tri.zip

by Kiszewski et al. (2004).[5] Altitude data are taken from the CGIAR's SRTM30 dataset.[6]. Means of precipitation, temperature, and suitabilities for specific crops are taken from the FAO-GAEZ data portal.[7] Similar suitability measures have been used by Alesina et al. (2013) and Alsan (2015). Correlations in rainfall are computed using the Matsuura and Willmott (2007) gridded series.[8] We join each market to the nearest point in these data and compute correlations in annual rainfall over the period 1900-2000. Humidity data are taken from the Climatic Research Unit at the University of East Anglia.[9]

For the variables that require geographic information systems data (that is, the coastal and river indicators, as well as those using raster data), we begin with a district map for modern India.[10] We compute the coastal and river indicators at this level, and compute other geographic variables by averaging over raster points within a district. If a market in our data shares the name of a modern-day district (or an updated name, as in the case of Benares and Varanasi), we make a unique $1:1$ merge between the market and the modern district polygon. Otherwise, we make a $1:m$ merge of all districts that split from the erstwhile district that previously shared the name of the market.

2.3. **Summary statistics.** Summary statistics are presented in Table 1. Some general patterns are apparent from this table. First, relative to a maximum number of observations of $\frac{206^2-206}{2} = 21,115$, we typically have fewer pairwise correlation coefficients. This is because not all products are traded in all markets. Second, while the degree of price integration is relatively high (> 0.8 for both wheat and rice), there is variation in price integration both across space and across markets. Some market pairs exhibit negative price correlations. Market integration is more limited for salt than for rice and wheat; the average price correlation for salt (< 0.35) is lower, and more than a quarter of these correlations are negative. One possible explanation of this lower correlation is the limited number of inland production sites for salt, which limit arbitrage opportunities in response to shocks, causing low average salt price correlations across markets. Linguistic distances range from close to 0 (i.e. market pairs in which both markets are dominated by the same language) to 1 (i.e. market pairs in which the dominant languages spoken are unrelated).

[5]We are grateful to Marcella Alsan for providing us with these data.
[6]`http://www.diva-gis.org/gdata`
[7]`http://www.fao.org/nr/gaez/en/`.
[8]`http://climate.geog.udel.edu/climate`
[9]`https://crudata.uea.ac.uk/cru/data/hrg/tmc/grid_10min_reh.dat.gz`
[10]In particular, we use the boundaries reported by `www.gadm.org`.

TABLE 1. Summary statistics

	(1) Mean	(2) s.d.	(3) Min	(4) Max	(5) N
Correlation: Wheat	0.81	0.22	-1	1	15,652
Correlation: Salt	0.33	0.53	-1	1	20,683
Correlation: Rice	0.81	0.16	-0.25	1	20,909
Linguistic Distance	0.42	0.39	0.000061	1.00	21,115
Genetic Distance	0.0026	0.0016	1.8e-07	0.010	21,115
Ln Distance in KM	6.85	0.71	1.99	8.24	21,115
Same Province	0.11	0.32	0	1	21,115

3. RESULTS

3.1. **Results by market.** Before presenting estimates of (1), we present preliminary descriptive evidence. For each market i in our data, we estimate:

$$\rho_{ij}^{p} = \beta_i^{p} LinguisticDistance_{ij} + x_{ij}^{p\prime}\gamma^{p} + \epsilon_{ij}^{p}. \tag{4}$$

In (4), ρ_{ij}^{p} and x_{ij}^{p} are defined as in (1). For each market i, we obtain a coefficient β_i^{p} that captures the degree to which its prices more closely track prices at other markets that are more linguistically similar, conditional on other measures of distance and dissimilarity.

To present these results, we order markets from those with the most negative estimates of β_i^{p} to those with the most positive estimates and present the point estimates and 95% confidence intervals in Figures 3, 4, and 5. For each of the three major crops, the majority of coefficients is negative and significant. This demonstrates two points. First, our main results pooling together all market pairs are not driven by a small number of markets. Second, (1) yields estimates of β^{p} that capture a central tendency in the sample.

3.2. **Main results.** In Table 2, we present our main estimates of (1). Across the three major crops, linguistic distance predicts reduced market integration. This is statistically significant in all specifications save one: wheat with controls but without fixed effects. There are several ways to consider the magnitudes involved. First, taking the estimates from column (4), a one standard deviation increase in linguistic distance, conditional on controls and fixed effects, predicts a reduction in the price correlation between markets i and j by 0.121 standard deviations for wheat, 0.167 standard deviations for salt, and 0.088 standard deviations for rice.

An alternative approach to magnitudes is to divide $\hat{\beta}^{p}$ by the coefficient estimated on $\ln(Distance)$ in column (4). This suggests that moving one unit in linguistic distance (i.e. from a closely-related language to an unrelated one) predicts a reduction in the price correlation comparable to a distance change of 789% for wheat, 8,200% for salt, and

FIGURE 3. Results by market: Wheat

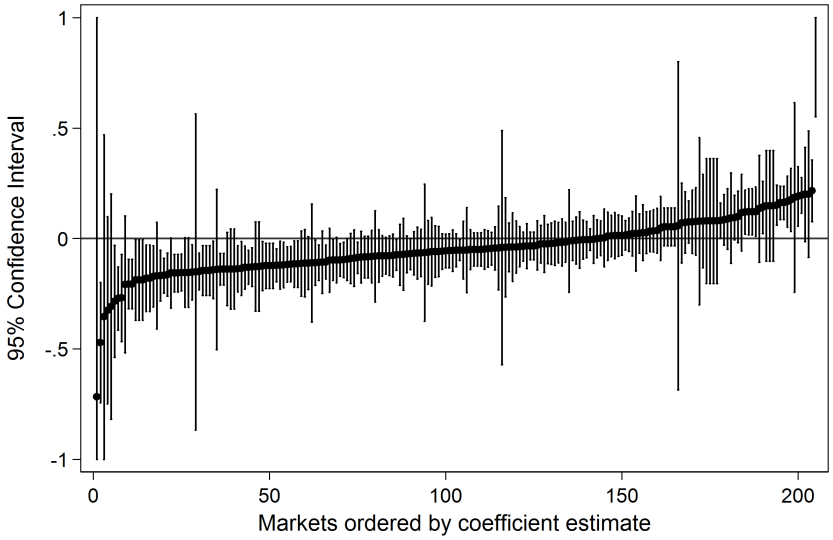
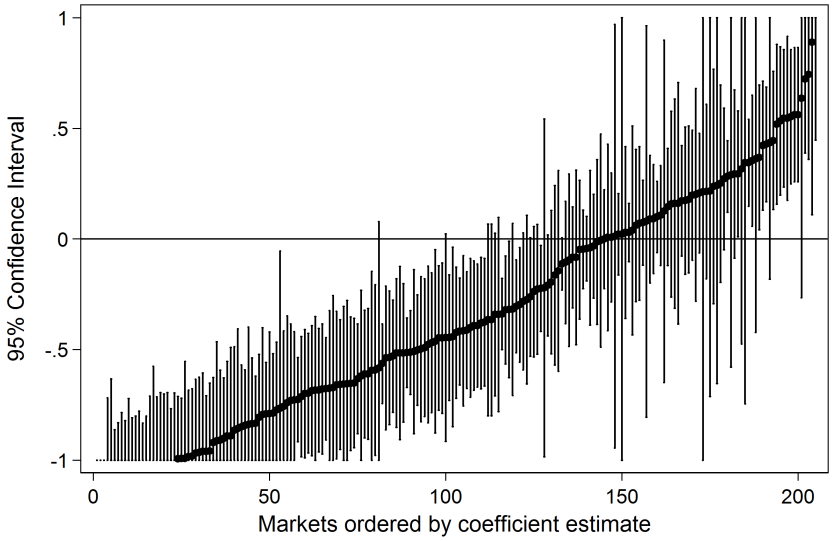


FIGURE 4. Results by market: Salt



210% for rice. These large numbers are driven in part by the small coefficients estimated on distance once additional controls are included.

Another way to think about these magnitudes is to consider the three markets of Ludhiana, Dacca, and Madurai. Column (4) gives estimates of $\hat{\beta}$ of -0.0667 for wheat, -0.224 for salt, and -0.035 for rice. The distance of Punjabi and Bengali with $\delta = 0.05$ is 0.077, which is similar to the linguistic distance of 0.081 between Ludhiana and Dacca. Similarly, the distance of Punjabi and Tamil with $\delta = 0.05$ is 1, corresponding closely to the

FIGURE 5. Results by market: Rice

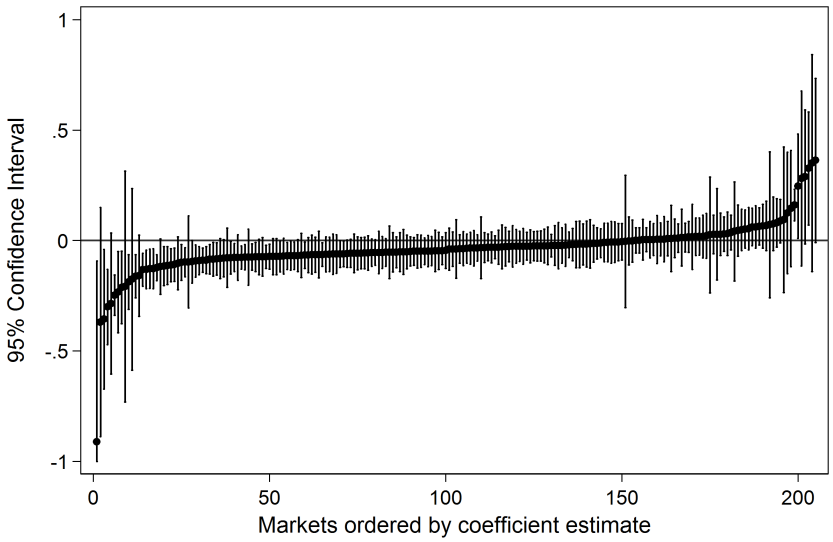


TABLE 2. Main results

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
N	15,652	15,652	15,652	15,652
		Correlation: Salt		
Linguistic Distance	-0.494***	-0.555***	-0.422***	-0.224***
	(0.074)	(0.085)	(0.071)	(0.073)
N	20,683	20,683	20,683	20,683
		Correlation: Rice		
Linguistic Distance	-0.083***	-0.073***	-0.056***	-0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

distance of $.98$ between Ludhiana and Madurai. Linguistic distances between Ludhiana and Dacca lower price correlations by $\hat{\beta}^p \times LinguisticDistance_{ij}$, or $.005$ for wheat, $.018$ for salt, and $.003$ for rice. Between Ludhiana and Madurai, the predicted reduction in price correlation is much larger, corresponding roughly to $\hat{\beta}^p$.

For comparison with other studies, Lameli et al. (2015) report an elasticity of inter-regional German trade with respect to dialectical similarity of 0.253. Falck et al. (2012)

find that a one standard deviation increase in dialect similarity within Germany increases gross migration flows between a region pair by about 6%. Melitz (2008) finds that sharing a common language raises trade within a country pair by 0.97 log points.

4. MECHANISMS

In this section, we outline the mechanisms suggested in both the economic and historical literatures that provide plausible links between genetic distance and market integration. We then assess these empirically to the extent our data allow.

4.1. **Mechanisms in the literature.** A recent economic literature has emphasized several possible channels that might link linguistic distance to market outcomes, and several of these mechanisms are reflected in observations made about colonial Indian markets in the secondary historical literature. One branch of this economic literature has focused on the importance of barriers to the transmission of the traits that are transmitted across generations in driving dissimilarities in economic outcomes across populations. Spolaore and Wacziarg (2009, 2016) have shown that, at the level of countries, a genetic distance measure of the time since the most recent common ancestor predicts greater dissimilarity in levels of economic development, and interpret this as a proxy for barriers to the diffusion of "implicit beliefs, customs, habits, biases, [and] conventions," among other traits.

Alternatively, differences in language may proxy for differences in tastes. Atkin (2013, 2016) has shown that differences in regional tastes in India shape what consumers are willing to buy and have consequences for price movements upon liberalization. Where the local market for a good is thin due to these differences in taste, we might anticipate prices that do not track those in other South Asian markets.

Other branches of the economic literature suggest mechanisms by which language barriers may inhibit market integration by raising trade costs. For example, linguistic distance may affect the costs of acquiring information; Gomes (2014) shows that Africans who do not share the language of the community experience worse health outcomes because they face difficulties in acquiring health information. Allen (2014) finds that ethnic differences predict greater trade costs across markets in the Philippines today. Alternatively, linguistic distance may act as a barrier to flows of people, via costs of migration, establishing business connections, or through xenophobia. Bai and Kung (2014) find that genetic distance inhibited technological transfer in historical China via reduced interaction. Falck et al. (2012) demonstrate that dialects act as barriers to migration in Germany, while Lameli et al. (2015) stress costs of doing business in explaining a similar result for trade flows. Rauch and Trindade (2002) show that ethnic Chinese networks facilitate international trade.

This branch of the economics literature aligns most closely with descriptions of trade in the secondary literature on Indian history. Collins (1999) cites linguistic barriers as an explanation of the low migration rates in India and hence as a limiting factor on price integration. Several writers have highlighted the importance of trade networks that corresponded with linguistic divisions. In colonial India, trading networks were often caste or kinship networks (Bhattacharya, 1983; Kessinger, 1983). Markovits (2008, p.188-196) mentions several such "middlemen minorities," including the Marwaris, Gujaratis, Parsis, Sindhis, Chettiars, Khatris, Aroras, Multanis, Bhatias, Khojas, Lohanas, Bohras, Memons, Banias, Pathans, Vanis, Shravaks, Agarwals, Maheshwaris, Oswals, Khandelwals, and Porwals.[11] Divekar (1983) adds to this the Afghans, Voras, Lingayat Banjigs, Komtis, and Vanjaris. These groups, he argues, contributed to the "unification of markets in India." They adopted new forms of business partnership and circulated information over space. Kumar (1983) and McAlpin (1974), similarly, highlight the role of the Banjaras.

Linguistic distance may also make it more difficult to acquire a language in which trade is conducted or to acquire common levels of education; Isphording and Otten (2014), Jain (2015), Laitin and Ramachandran (2016), and Shastry (2012) all find evidence that the costs of acquiring a new language – or education provided in that new language – are higher for those whose mother tongue is more dissimilar to the new language. Finally, linguistic distance may proxy for differences in preferences over public goods, redistribution, and the provision of infrastructure. In a series of papers, Desmet et al. (2016a, 2012, 2015) have shown the importance of linguistic cleavages in driving political outcomes, including those affecting public goods. If these public goods and infrastructure investments affect trade costs, they may help explain our main result.

4.2. **Mechanisms: Evidence.** To evaluate whether linguistic distance operates as a proxy for a broader set of barriers to the transmission of information, technology, and culture, we compute a measure of the genetic distance between the markets in our data. We show that, while linguistic distance and genetic distance are strongly correlated, neither one is a sufficient statistic for the other. We further show that it is the highest-level distinctions in our data, such as between Indo-European and Dravidian languages that drive our results.

To test whether missing markets, for example due to differences in tastes, drive the correlation between linguistic distance and market integration, we evaluate whether linguistic distance predicts whether the price of a good is reported at two markets in the same year, and we further evaluate whether markets that are more linguistically distant

[11] Roy (2014) similarly discusses the role of Marwaris, Banias, Parsis and Khojas.

from their neighbors experience more volatile prices. We find little evidence of missing markets for major crops increasing with linguistic distance. There is only limited evidence prices are more variable at markets that are more linguistically different from those around them.

To evaluate whether the presence of trading networks sharing a common tongue drives our result (for example, if small communities of traders have lower costs of establishing themselves in regions where the dominant language resembles their own), we correlate linguistic distance with the common presence of communities such as the Marwaris or Parsis. We find little evidence that the co-presence of these communities correlates with linguistic distance. In a related test for the costs of information, we examine whether linguistic distance correlates with differences in literacy rates. While linguistically distant markets have more dissimilar literacy rates, that this does not diminish the correlation of linguistic distance with market integration.

Finally, to examine whether linguistic distance proxies for shared preferences over public goods, in particular those that facilitate trade, we show that more linguistically distant markets spend less time both connected to the railway network, but that this does not fully account for our main result.

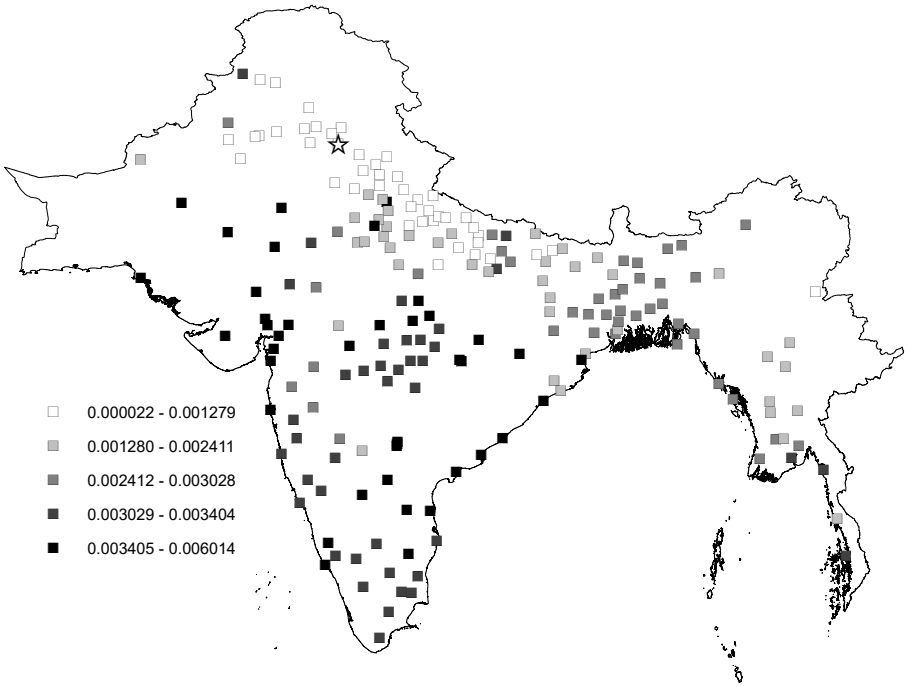
4.2.1. *Genetic distance.* We obtain data on genetic distance from Pemberton et al. (2013). Similar to the data used by Spolaore and Wacziarg (2009), these data contain pairwise Weir and Cockerham (1984) F_{ST} coefficients based on differences in allele frequencies from microsatellites. While the raw data report coefficients based on 5795 individuals from 267 human populations, we restrict ourselves to the data on ethnic groups indigenous to South Asia. These are the Balochi, Brahui, Burusho, Hazara, Kalash, Makrani, Pathan, Sindhi, Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Marathi, Marwari, Oriya, Parsi, Punjabi, Tamil, and Telugu. While these groups cover the majority of the population in our sample, there are some major missing groups, of which Urdu is the largest.

Following Spolaore and Wacziarg (2009), abusing notation slightly, and given population shares of groups i and j in districts 1 and 2 of s_{1i} and s_{2j} with genetic distance F_{ST}^{ij}, we compute genetic distance between districts as:

$$GD_{12} = \sum_i \sum_j (s_{1i} \times s_{2j} \times F_{ST}^{ij}). \tag{5}$$

Note that we re-scale s_{1i} and s_{2j} as fractions of the population matched to the genetic data, rather than as fractions of the full district population. We present a map of genetic distances from Ludhiana in Figure 6. This has many similarities to Figure 2.

FIGURE 6. Ludhiana: Genetic distances



Other regions of South Asia that are proximate to the Punjab are more genetically similar, though it is clear that South Indian groups in Dravidian-speaking regions are more genetically dissimilar, conditional on physical distance. The apparent proximity with Burma is overstated due to the lack of coverage of major Burmese populations in the genetic data.

Our aim is to assess whether linguistic distance proxies for broader (and possibly deeper) barriers to the diffusion of information, culture, and technology. We re-estimate (1), first with genetic distance as an outcome, and second with genetic distance as an additional control. We report results in Table 3. Linguistic and genetic distance are correlated, even conditional on our baseline fixed effects and controls. Genetic distance itself predicts less market integration and diminishes the coefficient on linguistic distance, but does not fully eliminate it in any specifications where linguistic distance was significant in Table 2.

4.2.2. *Coarse and fine distinctions.* Recall that, in our baseline analyses, we computed the distance between any two languages i and j as:

$$d_{ij} = 1 - \left(\frac{SharedBranches}{15}\right)^{\delta}$$

TABLE 3. Genetic Distance

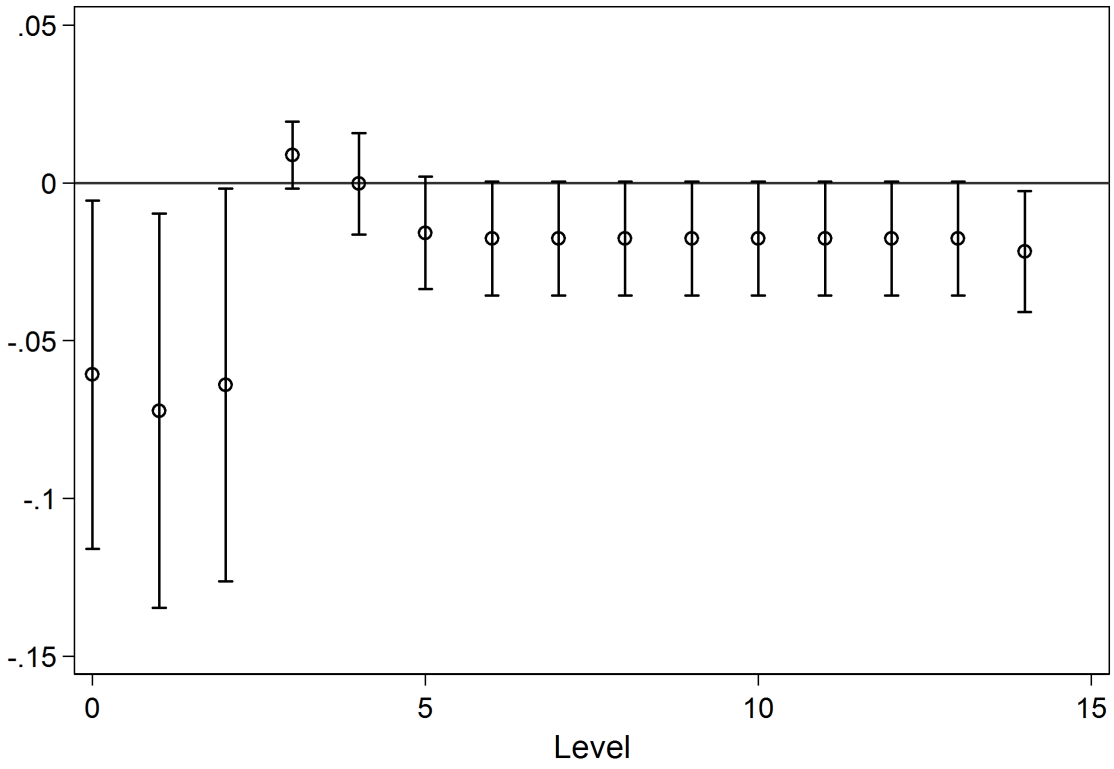
	(1)	(2)	(3)	(4)
		Genetic Distance X 100		
Linguistic Distance	0.046***	0.105***	0.041**	0.027**
	(0.014)	(0.012)	(0.020)	(0.013)
N	21,115	21,115	21,115	21,115
		Correlation: Wheat		
Linguistic Distance	-0.253***	-0.159***	-0.021	-0.062**
	(0.036)	(0.035)	(0.025)	(0.030)
Genetic Distance X 100	-0.063*	-0.283***	-0.036	-0.058**
	(0.036)	(0.050)	(0.025)	(0.026)
N	15,652	15,652	15,652	15,652
		Correlation: Salt		
Linguistic Distance	-0.463***	-0.474***	-0.400***	-0.224***
	(0.072)	(0.085)	(0.071)	(0.073)
Genetic Distance X 100	-0.661***	-0.774***	-0.451***	-0.009
	(0.146)	(0.167)	(0.129)	(0.144)
N	20,683	20,683	20,683	20,683
		Correlation: Rice		
Linguistic Distance	-0.076***	-0.057***	-0.051***	-0.034***
	(0.019)	(0.012)	(0.020)	(0.010)
Genetic Distance X 100	-0.167***	-0.154***	-0.113*	-0.034*
	(0.064)	(0.030)	(0.064)	(0.018)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

While this follows the convention in the literature, it does not allow us to distinguish whether coarser distinctions between e.g. Indo-European and Dravidian languages drive our results, or whether lesser divisions, as between Bengali and Punjabi, do so. We replace d_{ij} with a dummy for having $\leq N$ shared branches, for $N = \{1, ..., 15\}$. We re-estimate (1), and present our results in Figures 7, 8, and 9. These correspond to column (4) with fixed effects and controls. In all three figures, it is clear that coarser distinctions matter more than finer ones.

Consider a language such as Gujarati (Indo-European, Indo-Iranian, Indo-Aryan, Intermediate Divisions, Gujarati, Gujarati). It has no branches in common with a Dravidian language such as Tamil. It shares one branch with languages such as Yiddish that are Indo-European but not Indo-Iranian. It shares two branches with languages such as Balochi that are Indo-Iranian but not Indo-Aryan. It shares three branches with an Indo-Aryan language such as Hindi that is classified under "Western Hindi" rather than "Intermediate Divisions." It shares four branches with a language such as Nepali that is within these "Intermediate Divisions," but is not within the Gujarati sub-class. It shares five branches with other Gujarati languages such as Jandavra. In all three figures, language divisions with two common branches or fewer yield visibly greater differences

FIGURE 7. Results by level: Wheat



than finer distinctions, suggesting it is divisions on the scale of Gujarati-Tamil, Gujarati-Yiddish, and Gujarati-Balochi that drive our results, rather than finer distinctions as between Gujarati and Hindi, Nepali, or Jandavra.

4.2.3. *Missing markets.* To assess the degree to which our results are driven by differences in tastes that can lead to thin or missing markets across linguistic divides, we take two approaches. First, we test whether linguistic distance predicts how frequently prices are available for two markets in the same year. Taking N_{ij}^p as the number of common price observations at markets i and j for product p, we estimate (1), except that we now take N_{ij}^p as the dependent variable, and no longer control for minimum year, maximum year or the number of common observations. Results are presented in Table 4. There is only weak evidence of missing markets correlating with genetic distance; while we find a negative correlation between linguistic distance and N_{ij}^p, no such correlation is available for salt or rice. Although we do not report these here, we find similar failures of linguistic distance to predict N_{ij}^p when using lesser crops from the data such as barley and maize.

FIGURE 8. Results by level: Salt

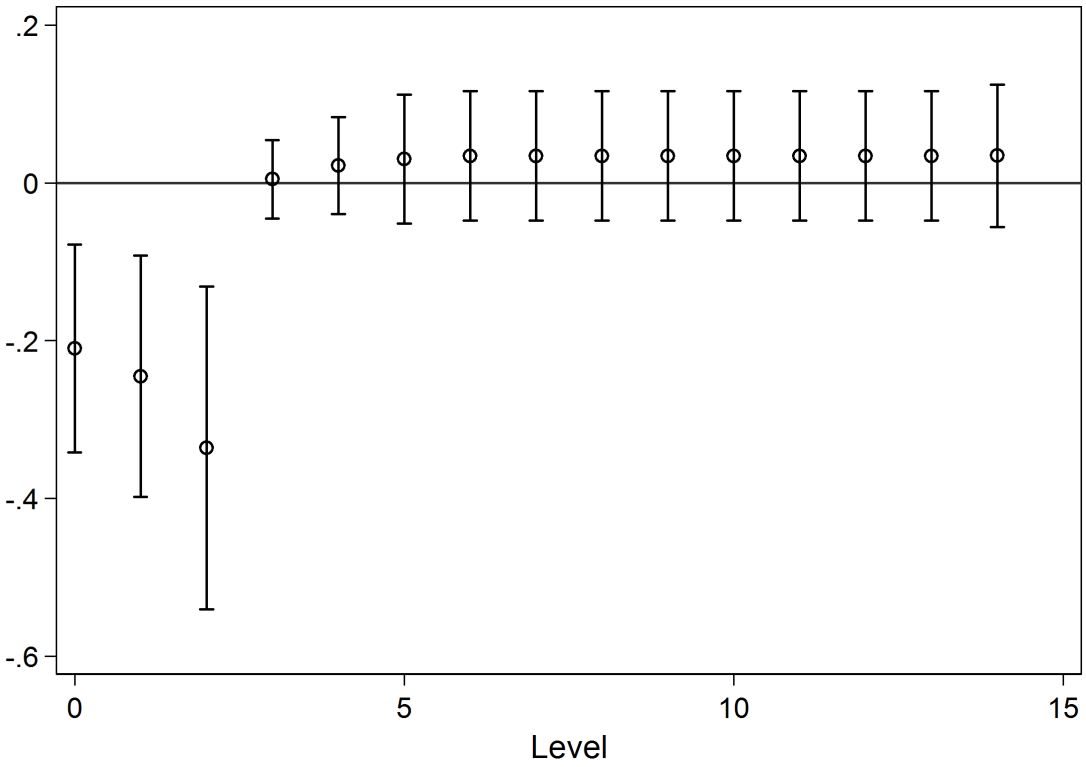
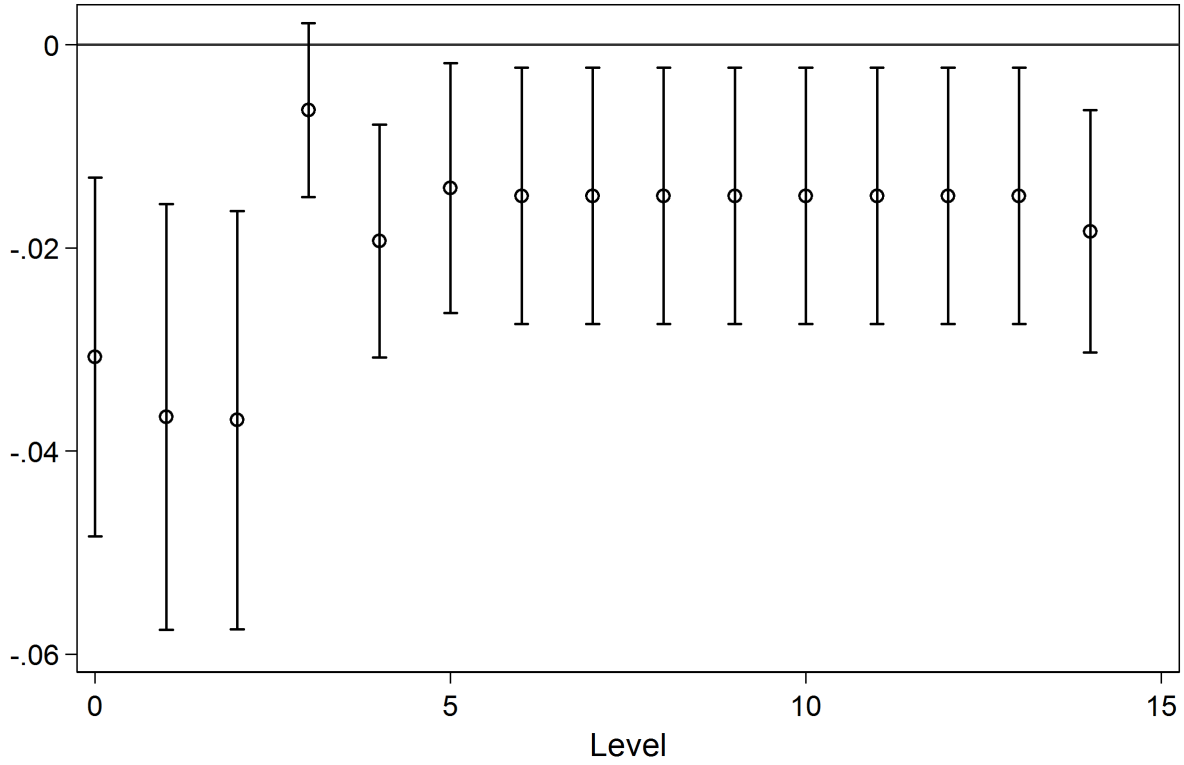


TABLE 4. Missing markets: Number of common years

	(1)	(2)	(3)	(4)
		Observations: Wheat		
Linguistic Distance	-37.518***	-15.483***	-36.672***	-13.412***
	(2.450)	(2.325)	(3.045)	(2.183)
N	21,115	21,115	21,115	21,115
		Observations: Salt		
Linguistic Distance	-1.334	-0.014	-2.666	-0.018
	(1.259)	(0.072)	(1.753)	(0.156)
N	21,115	21,115	21,115	21,115
		Observations: Rice		
Linguistic Distance	-1.441	0.011	-3.126	-0.097
	(1.316)	(0.085)	(1.938)	(0.165)
N	21,115	21,115	21,115	21,115
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

FIGURE 9. Results by level: Rice



As a second approach, we evaluate whether markets that are more linguistically distant than those within a set radius experience prices that are more volatile. For each market i, we keep the other markets within 500km and take the average of their linguistic distance from i (denoted $\bar{Linguistic Distance}_{ij}$) as well as the average of the controls (denoted $\bar{x}^p_{ij}$). We estimate:

$$CV_i^p = \beta^p \bar{Linguistic Distance}_{ij} + \bar{x}^{p\prime}_{ij}\gamma^p + \epsilon_i^p. \tag{6}$$

In (6), CV_i^p is the coefficient of variation of the price of product p at market i. We estimate (6) by OLS and report robust standard errors Results are presented in Table 5. While we find evidence that wheat prices are more volatile at markets that are more linguistically distant from others in their neighborhood, we find no similar evidence for rice or salt.

4.2.4. *Trading communities.* In order to identify whether our results are driven by the co-location of communities of traders who served as information networks, we focus on one group that has received particular attention in the literature: the Marwaris. By

TABLE 5. Missing markets: Volatility

	(1) *CV: Whe*	(2) *CV: Sal*	(3) *CV: Ric*
Linguistic Distance	0.127***	0.024	-0.113
	(0.049)	(0.058)	(0.279)
N	178	204	205
Fixed Effects	No	No	No
Controls	Yes	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Robust standard errors in parentheses. All regressions are OLS and include a constant. Controls are averages of minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. .

1920, there were between than 200,000 and 400,000 Marwaris outside of the Rajputana Agency, working mostly as traders (Markovits, 2008). These traders drew on capital and personnel from throughout the subcontinent. They gained dominant positions in regional trade, importing, exporting and moneylending. These communities held assets jointly in patrilineal extended families, sharing information and personnel (Roy, 2014).

For each pair of markets i and j, we estimate the absolute difference in Marwari share, or $AD_{ij}^{Marwari} = |s_i^{Marwari} - s_j^{Marwari}|$. We then estimate (1) with $AD_{ij}^{Marwari}$ as both an outcome and as a control. That is: we test whether linguistic distance predicts the co-location of Marwaris across district pairs, and the degree to which the co-presence of this trading community can account for the conditional correlation between lingusitic distance and market integration. Results are presented in Table 6. There is little evidence of linguistic distance driving distances in the presence of this trading community, and little evidence that it explains price integration.[12]

4.2.5. *Literacy.* To evaluate whether linguistic distance acts as a barrier to the transmission of information via the acquisition of a common means of communication, we test whether differences in literacy rates across markets correlate with linguistic distance, and whether the correlation between linguistic distance and market integration is diminished by controlling for dissimilarities in literacy. For data on literacy, we use the 1921 Census of India. These data report literacy at the district level, and we match each market to the district that contains it. As with the presence of trading communities, for each community, we take this difference as both an outcome and as a control.

[12]Results are similar if we perform the same exercise for the Parsis or English. Our results are unlikely to be explained by the spread of the English language: fewer than one tenth of one percent of the population in the 1901 census is recorded as "English" by language.

TABLE 6. Trading communities

	(1)	(2)	(3)	(4)
		Absolute difference in Marwaris share		
Linguistic Distance	-0.025**	0.001	0.055**	-0.001
	(0.012)	(0.001)	(0.024)	(0.001)
N	21,115	21,115	21,115	21,115
		Correlation: Wheat		
Linguistic Distance	-0.255***	-0.210***	-0.023	-0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
Abs. diff. in Marwaris share	0.066***	0.021*	-0.003	0.030*
	(0.014)	(0.011)	(0.016)	(0.017)
N	15,652	15,652	15,652	15,652
		Correlation: Salt		
Linguistic Distance	-0.506***	-0.555***	-0.414***	-0.224***
	(0.073)	(0.085)	(0.071)	(0.073)
Abs. diff. in Marwaris share	-0.456***	-0.170	-0.158*	0.011
	(0.099)	(0.163)	(0.096)	(0.164)
N	20,683	20,683	20,683	20,683
		Correlation: Rice		
Linguistic Distance	-0.085***	-0.073***	-0.054***	-0.035***
	(0.017)	(0.010)	(0.017)	(0.010)
Abs. diff. in Marwaris share	-0.074	-0.040***	-0.028	0.015
	(0.065)	(0.012)	(0.073)	(0.013)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

We present results in Table 7. More linguistically distant markets have more dissimilar literacy rates, but this does little to predict price correlations or explain away their correlation with linguistic distance.

4.2.6. *Infrastructure.* In order to evaluate whether linguistic proximity predicts common preferences for public goods that may facilitate trade, we test whether linguistic distance reduces the degree to which market pairs were both connected to the colonial railway network. Using the 1934 edition of *History of Indian Railways Constructed and In Progress*, we identify the year each market became connected to the colonial railway. This source divides the Indian railway system into segments (e.g. "Karimganj to Badarpur") with a date of opening (in this example, 4-12-96) and length in miles (in this example, 12.00). We use these data to code the first date at which the district containing each market was connected to the Indian Railway system. For each market pair ij, we can then identify the number of years up to 1921 that both markets were connected to the railway system. We then estimate (1) with this variable as both an outcome and as a control. We present results in Table 8. More linguistically distant markets spend more time both connected to the railroad, but this does little to predict price correlations or

TABLE 7. Literacy Rate

	(1)	(2)	(3)	(4)
		Difference in Literacy 1921		
Linguistic Distance	10.432***	6.920***	6.826***	4.691***
	(1.505)	(1.869)	(1.086)	(1.264)
N	20,503	20,503	20,503	20,503
		Correlation: Wheat		
Linguistic Distance	-0.247***	-0.206***	-0.018	-0.067**
	(0.034)	(0.035)	(0.025)	(0.030)
Difference in Literacy 1921	-0.001	-0.001	-0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.001)
N	15,125	15,125	15,125	15,125
		Correlation: Salt		
Linguistic Distance	-0.459***	-0.526***	-0.365***	-0.206***
	(0.071)	(0.084)	(0.069)	(0.072)
Difference in Literacy 1921	-0.006***	-0.007**	-0.007***	-0.005**
	(0.002)	(0.003)	(0.002)	(0.002)
N	20,077	20,077	20,077	20,077
		Correlation: Rice		
Linguistic Distance	-0.019	-0.064***	-0.017	-0.032***
	(0.025)	(0.008)	(0.025)	(0.010)
Difference in Literacy 1921	-0.006***	-0.001***	-0.006**	-0.001
	(0.002)	(0.000)	(0.002)	(0.001)
N	20,300	20,300	20,300	20,300
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

explain away their correlation with linguistic distance. One contributing factor to these results is the nature of the Indian railways, which were often built to track pre-existing trade routes (Andrabi and Kuehlwein, 2010).

5. ROBUSTNESS

5.1. **Main Robustness.** In this section, we demonstrate the robustness of our results to selection on unobservables, show that we can obtain similar results with other crops and with wages, and we discuss a number of additional exercises that we present in the appendix.

5.1.1. *Selection on unobservables.* To demonstrate robustness to selection on unobservables, we use the approach of Altonji et al. (2005) as implemented by Bellows and Miguel (2009) and Nunn and Wantchekon (2011). We estimate (1) with either a limited set of controls or with a full set of controls, and compute:

$$AET = \frac{\beta^{FullControls}}{\beta^{RestrictedControls} - \beta^{FullControls}} \tag{7}$$

TABLE 8. Railway connections

	(1)	(2)	(3)	(4)
		Years Both Connected to Railroad		
Linguistic Distance	-4.388**	-0.852*	-4.349*	-0.236
	(2.138)	(0.485)	(2.406)	(0.452)
N	21,115	21,115	21,115	21,115
		Correlation: Wheat		
Linguistic Distance	-0.258***	-0.210***	-0.026	-0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
Years Both Connected to Railroad	-0.000	0.000	-0.000	-0.000
	(0.001)	(0.000)	(0.000)	(0.000)
N	15,652	15,652	15,652	15,652
		Correlation: Salt		
Linguistic Distance	-0.493***	-0.555***	-0.423***	-0.224***
	(0.074)	(0.085)	(0.071)	(0.073)
Years Both Connected to Railroad	0.000	0.000	-0.000	-0.000
	(0.001)	(0.000)	(0.001)	(0.000)
N	20,683	20,683	20,683	20,683
		Correlation: Rice		
Linguistic Distance	-0.081***	-0.073***	-0.055***	-0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
Years Both Connected to Railroad	0.001	0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

We report result where the restricted set of controls is either empty or contains only $\ln(Distance)$. Larger values of this statistic imply that the selection on unobservables would need to have a larger effect on β relative to that of observables in order to be consistent with a true β of 0. Results are presented in Table 9. The coefficient estimates for wheat are sensitive to controls regardless of what is in the base set of controls, but are not as sensitive to the addition of fixed effects. Other results initially appear sensitive to adding fixed effects and controls together, but this is driven by $\ln(Distance)$. Once this is included as a baseline control, AET is negative (i.e. controls push β away from zero) or greater than one.

5.1.2. *Other crops.* Although we have focused our analysis on the crops whose prices are reported most in the data (wheat, salt and rice), we are able to show similar results for a wide range of other crops. These data are again taken from *Wages and Prices in India*. We present estimates of (1) for these other prices and wages in Tables 10, 11, 12, and 13. Several other prices show patterns similar to our main results. Where the conditional correlation between market integration and linguistic distance is insignificant,

TABLE 9. Altonji-Elder-Taber Statistics

		Correlation: Wheat	
Baseline: No Controls	4.476	0.0977	0.351
Baseline: ln(distance)	2.437	0.141	0.562
		Correlation: Salt	
Baseline: No Controls	-9.171	5.866	0.827
Baseline: ln(distance)	-17.26	-4.958	1.973
		Correlation: Rice	
Baseline: No Controls	7.219	2.051	0.714
Baseline: ln(distance)	1.495	-2.525	-39.48
Fixed Effects	Yes	No	Yes
Controls	No	Yes	Yes

TABLE 10. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Arhar Dal		
Linguistic Distance	-0.498***	-0.287***	-0.304***	-0.123***
	(0.083)	(0.053)	(0.069)	(0.048)
N	11,466	11,466	11,466	11,466
		Correlation: Bajra		
Linguistic Distance	-0.340***	-0.438***	-0.093**	-0.156***
	(0.030)	(0.043)	(0.039)	(0.043)
N	6,091	6,091	6,091	6,091
		Correlation: Barley		
Linguistic Distance	-0.237*	-0.357**	-0.216**	-0.092
	(0.133)	(0.175)	(0.085)	(0.099)
N	5,465	5,465	5,465	5,465
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

this is often for products whose pairwise price correlations we can compute for a much smaller set of market pairs than our main results.

5.2. **Additional robustness.** In the remainder of this section, we briefly outline robustness checks that are reported in the appendix.

5.2.1. *Sample.* In Table A1, we restrict our sample to modern India. In Table A2, we remove any negative price correlations from the sample. In Table A3, we remove outliers by discarding the top and bottom 5% of observations by values of ρ_{ij}^{p}. In Table A4, we instead remove outliers by discarding the top and bottom 5% of observations by values of linguistic distance. Tables A5 and A6 report results using only price observations from

TABLE 11. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Gram		
Linguistic Distance	-0.204***	-0.102***	-0.149***	-0.053**
	(0.034)	(0.014)	(0.022)	(0.022)
N	16,470	16,470	16,470	16,470
		Correlation: Jawar		
Linguistic Distance	-0.353***	-0.434***	-0.068	-0.209***
	(0.041)	(0.030)	(0.048)	(0.033)
N	7,985	7,985	7,985	7,985
		Correlation: Kangni		
Linguistic Distance	-0.520	-0.004	-0.799*	0.218
	(0.714)	(0.337)	(0.469)	(0.283)
N	1,275	1,275	1,275	1,275
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE 12. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Maize		
Linguistic Distance	-0.503***	-0.285***	0.009	-0.003
	(0.049)	(0.059)	(0.079)	(0.052)
N	2,850	2,850	2,850	2,850
		Correlation: Marua		
Linguistic Distance	-0.054	-0.139***	0.034	0.002
	(0.043)	(0.030)	(0.028)	(0.025)
N	1,275	1,275	1,275	1,275
		Correlation: Bulrush Millet		
Linguistic Distance	-0.146***	-0.162***	-0.057**	-0.090***
	(0.024)	(0.035)	(0.023)	(0.029)
N	6,322	6,322	6,322	6,322
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

before or after 1891 (the midpoint in the sample) to compute ρ_{ij}^{p}. Across these sample restriction exercises, results remain similar to the baseline.

TABLE 13. Other crops

	(1)	(2)	(3)	(4)
		Correlation: Great Millet		
Linguistic Distance	-0.145***	-0.116***	0.014	-0.045***
	(0.033)	(0.015)	(0.017)	(0.016)
N	8,368	8,368	8,368	8,368
		Correlation: Lesser Millet		
Linguistic Distance	-0.520***	-0.533***	-0.264***	-0.225***
	(0.125)	(0.103)	(0.102)	(0.085)
N	253	253	253	253
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

5.2.2. *Measures of linguistic distance and market integration.* In Table A7 we replace our baseline measure of market integration with the natural logarithm of the correlation coefficient. Similarly, in Table A8 we replace our main measure with centiles of the correlation coefficient. In Table A9 we replace our baseline measure of linguistic distance with an alternative in which $\delta = 0.5$. In Table A10, we instead use the pairwise distance between the largest language in each district to compute linguistic distance. Both exercises give results similar to those in Table 2.

5.2.3. *Standard errors.* Tables A11 and A12 present alternative approaches to standard errors. Rather than clustering by market i and market j, we report two-way clustering by either the largest language in each district or by the province in which each district falls.

5.2.4. *Cost distance.* While in our baseline we control for the natural logarithm of pairwise distance in km, we can show that our results survive controlling for an alternative cost distance measure constructed by Özak (2010, 2013). Results appear in Table A14 and are almost unchanged.

5.2.5. *Convergence.* Because it is possible that the gradual erosion of a large price gap across two markets could produce a negative correlation in the prices recorded in the two markets, we show that our results survive controlling for the mean absolute log price difference between any two markets. Results are presented in Table A13 and the results are little different from our main results.

6. CONCLUSION

In this paper, we have shown that, conditional on several measures of distance and dissimilarity, markets in colonial South Asia that were more linguistically distant from each other displayed less market integration. This holds across multiple products and markets, and survives several sensitivity checks. Genetic distance, literacy gaps, and lack of railway connections may help explain these results, but are not sufficient statistics for them. There is less evidence for missing markets and presence of trading communities as mechanisms. We have shown, then, that cultural and linguistic barriers are salient to the functioning of markets and that their importance is not limited to political economy or post-colonial, modern economies. Furthermore, the contribution of these cultural factors to market integration, or lack thereof, are substantial relative to other factors hindering market integration such as physical distance. More linguistically-similar markets are more likely to be connected earlier via transport infrastructure (the colonial railway system), but this connection alone does not explain away the effect. These results indicate the importance and persistence of cultural differences for market integration, trade, and price volatility.

REFERENCES

Adams, J. and West, R. C. (1979). Money, prices, and economic development in India, 1861–1895. *The Journal of Economic History*, 39(01):55–68.

Alesina, A., Giuliano, P., and Nunn, N. (2013). On the Origins of Gender Roles: Women and the Plough. *The Quarterly Journal of Economics*, 128(2):469–530.

Allen, R. C. (2007). India in the great divergence. *The new comparative economic history: Essays in honor of Jeffrey G. Williamson*, pages 9–32.

Allen, T. (2014). Information frictions in trade. *Econometrica*, 82(6):2041–2083.

Alsan, M. (2015). The effect of the tsetse fly on African development. *The American Economic Review*, 105(1):382–410.

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy*, 113(1):151–184.

Andrabi, T. and Kuehlwein, M. (2010). Railways and price convergence in British India. *The Journal of Economic History*, 70(02):351–377.

Ashraf, Q. and Galor, O. (2011). Dynamics and stagnation in the Malthusian epoch. *The American Economic Review*, 101(5):2003–2041.

Ashraf, Q. and Galor, O. (2013). Genetic diversity and the origins of cultural fragmentation. *The American Economic Review*, 103(3):528–533.

Atkin, D. (2013). Trade, tastes, and nutrition in India. *The American Economic Review*, 103(5):1629–1663.

Atkin, D. (2016). The caloric costs of culture: Evidence from Indian migrants. *The American Economic Review*, 106(4):1144–1181.

Bai, Y. and Kung, J. (2014). Does Genetic Distance have a Barrier Effect on Technology Diffusion? Evidence from Historical China. *Working Paper: Hong Kong University of Science and Technology*.

Bellows, J. and Miguel, E. (2009). War and local collective action in Sierra Leone. *Journal of Public Economics*, 93(11):1144–1157.

Bhattacharya, S. (1983). Regional Economy (1757–1857): Eastern India. *The Cambridge Economic History of India*, 2:270–95.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.

Chandavarkar, A. G. (1983). Money and credit, 1858–1947. *The Cambridge economic history of India*, 2:762–803.

Collins, W. J. (1999). Labor mobility, market integration, and wage convergence in late 19th century India. *Explorations in Economic History*, 36(3):246–277.

Derbyshire, I. D. (1987). Economic Change and the Railways in North India, 1860–1914. *Modern Asian Studies*, 21(03):521–545.

Desmet, K., Gomes, J., and Ortuño, I. (2016a). The geography of linguistic diversity and the provision of public goods. *CEPR Discussion Paper DP116*.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2012). The political economy of linguistic cleavages. *Journal of Development Economics*, 97(2):322–338.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2015). Culture, ethnicity and diversity. *NBER Working Paper No. 20989*.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2016b). Linguistic Cleavages and Economic Development. In *The Palgrave Handbook of Economics and Language*, pages 425–446. Springer.

Dickens, A. (2016). Ethnolinguistic Favouritism in African Politics. *Working Paper: York University*.

Divekar, V. (1983). Regional Economy (1757–1857): Western India. *The Cambridge Economic History of India*, 2:332–51.

Esteban, J., Mayoral, L., and Ray, D. (2012). Ethnicity and conflict: An empirical study. *The American Economic Review*, 102(4):1310–1342.

Estevadeordal, A., Frantz, B., Taylor, A. M., et al. (2003). The Rise and Fall of World Trade, 1870–1939. *The Quarterly Journal of Economics*, 118(2):359–407.

Falck, O., Heblich, S., Lameli, A., and Südekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, 72(2):225–239.

Federico, G. (2011). When did European markets integrate? *European Review of Economic History*, 15(1):93–126.

Frawley, W. J. (2003). *International encyclopedia of linguistics*, volume 4. Oxford university press.

Giuliano, P., Spilimbergo, A., and Tonon, G. (2014). Genetic distance, transportation costs, and trade. *Journal of Economic Geography*, 14(1):179–198.

Gomes, J. F. (2014). The health costs of ethnic distance: evidence from Sub-Saharan Africa. *ISER Working Paper Series 2014-33*.

Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural biases in economic exchange? *The Quarterly Journal of Economics*, 124(3):1095–1131.

Hurd, J. (1975). Railways and the Expansion of Markets in India, 1861–1921. *Explorations in Economic History*, 12(3):263–288.

Hutchinson, W. K. (2005). "Linguistic distance" as a determinant of bilateral trade. *Southern Economic Journal*, 72(1):1–15.

Isphording, I. E. and Otten, S. (2014). Linguistic barriers in the destination language acquisition of immigrants. *Journal of Economic Behavior & Organization*, 105:30–50.

Jacks, D. S., Meissner, C. M., and Novy, D. (2008). Trade Costs, 1870–2000. *The American Economic Review*, 98(2):529–534.

Jacks, D. S., O'Rourke, K. H., and Williamson, J. G. (2011). Commodity price volatility and world market integration since 1700. *Review of Economics and Statistics*, 93(3):800–813.

Jain, T. (2015). Common tongue: The impact of language on educational outcomes. *Indian School of Business WP ISB-WP/103/2011*.

Jia, R. (2014). Weather shocks, sweet potatoes and peasant revolts in historical China. *The Economic Journal*, 124(575):92–118.

Kessinger, T. G. (1983). Regional Economy (1757–1857): North India. *The Cambridge Economic History of India*, 2:242–270.

Kiszewski, A., Mellinger, A., Spielman, A., Malaney, P., Sachs, S. E., and Sachs, J. (2004). A global index representing the stability of malaria transmission. *The American journal of tropical medicine and hygiene*, 70(5):486–498.

Kumar, D. (1983). Regional Economy (1757–1857): South India. *The Cambridge Economic History of India*, 2:352–375.

Laitin, D. and Ramachandran, R. (2016). Language Policy and Human Development. *American Political Science Review*, 110(3):457–480.

Lameli, A., Nitsch, V., Südekum, J., and Wolf, N. (2015). Same same but different: Dialects and trade. *German Economic Review*, 16(3):290–306.

Laval, G., Patin, E., and Rueda, V. (2016). Achieving the American Dream: Cultural Distance, Cultural Diversity and Economic Performance. *Oxford Economic and Social History Working Paper 140*.

Markovits, C. (2008). *Merchants, traders, entrepreneurs: Indian business in the colonial era*. Springer.

Matsuura, K. and Willmott, C. (2007). Terrestrial Air Temperature and Precipitation: 1900-2006 Gridded Monthly Time Series, Version 1.01. *University of Delaware*.

McAlpin, M. (1983). Price Movements and Economic Activity (1860-1947). *The Cambridge Economic History of India*, 2:878–904.

McAlpin, M. B. (1974). Railroads, Prices, and Peasant Rationality: India 1860–1900. *The Journal of Economic History*, 34(03):662–684.

Melitz, J. (2008). Language and foreign trade. *European Economic Review*, 52(4):667–699.

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity. *The American Economic Review*, 102(4):1508–1539.

Nunn, N. and Puga, D. (2012). Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics*, 94(1):20–36.

Nunn, N. and Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. *The American Economic Review*, 101(7):3221–3252.

O'Rourke, K. H. and Williamson, J. G. (2002). When did globalisation begin? *European Review of Economic History*, 6(1):23–50.

Özak, Ö. (2010). The Voyage of Homo-Economicus: Some Economic Measures of Distance. *Working Paper: Department of Economics, Southern Methodist University*.

Özak, Ö. (2013). Distance to the technological frontier and economic development. *Working Paper Available at SSRN 1989216*.

Pascali, L. (2016). The wind of change: Maritime technology, trade and economic development. *Forthcoming in the American Economic Review*.

Pemberton, T. J., DeGiorgio, M., and Rosenberg, N. A. (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. *G3: Genes, Genomes, Genetics*, 7(2):g3–113.

Persson, K. G. (1999). *Grain markets in Europe, 1500–1900: Integration and deregulation*, volume 7. Cambridge University Press.

Ramankutty, N., Foley, J. A., Norman, J., and McSweeney, K. (2002). The global distribution of cultivable lands: current patterns and sensitivity to possible climate change. *Global Ecology and biogeography*, 11(5):377–392.

Rauch, J. E. and Trindade, V. (2002). Ethnic Chinese networks in international trade. *Review of Economics and Statistics*, 84(1):116–130.

Roy, T. (2012). *India in the world economy: from antiquity to the present*. Cambridge University Press.

Roy, T. (2014). Trading Firms in Colonial India. *Business History Review*, 88(1):9–42.

Shastry, G. K. (2012). Human capital response to globalization education and information technology in India. *Journal of Human Resources*, 47(2):287–330.

Shiue, C. H. and Keller, W. (2007). Markets in China and Europe on the Eve of the Industrial Revolution. *The American Economic Review*, 97(4):1189–1216.

Spolaore, E. and Wacziarg, R. (2009). The diffusion of development. *The Quarterly Journal of Economics*, 124(2):469–529.

Spolaore, E. and Wacziarg, R. (2016). Ancestry and development: New evidence. *Working Paper, Department of Economics, Tufts University*.

Studer, R. (2008). India and the great divergence: Assessing the efficiency of grain markets in eighteenth-and nineteenth-century India. *Journal of Economic History*, 68(02):393–437.

Waldinger, M. (2014). The economic effects of long-term climate change: Evidence from the little ice age. *Working Paper: London School of Economics*.

Weir, B. S. and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

Appendix: Not for publication

APPENDIX A. ADDITIONAL ROBUSTNESS: TABLES

TABLE A1. Restrict sample to India

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.268***	-0.217***	-0.044*	-0.074**
	(0.033)	(0.038)	(0.026)	(0.032)
N	10,854	10,854	10,854	10,854
		Correlation: Salt		
Linguistic Distance	-0.454***	-0.555***	-0.296***	-0.225***
	(0.078)	(0.084)	(0.087)	(0.081)
N	12,880	12,880	12,880	12,880
		Correlation: Rice		
Linguistic Distance	-0.010	-0.053***	-0.000	-0.012**
	(0.014)	(0.006)	(0.019)	(0.006)
N	13,040	13,040	13,040	13,040
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A2. No negative correlations

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.243***	-0.207***	-0.028	-0.066**
	(0.031)	(0.035)	(0.024)	(0.029)
N	15,479	15,479	15,479	15,479
		Correlation: Salt		
Linguistic Distance	-0.345***	-0.228***	-0.263***	-0.077*
	(0.035)	(0.037)	(0.053)	(0.046)
N	14,230	14,230	14,230	14,230
		Correlation: Rice		
Linguistic Distance	-0.089***	-0.073***	-0.061***	-0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
N	20,768	20,768	20,768	20,768
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A3. Remove outliers by price correlation

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-0.191***	-0.178***	-0.020	-0.042
	(0.024)	(0.028)	(0.021)	(0.026)
N	14,243	14,243	14,243	14,243
			Correlation: Wheat	
Linguistic Distance	-0.230***	-0.204***	-0.035	-0.066**
	(0.038)	(0.035)	(0.025)	(0.030)
N	14,586	14,586	14,586	14,586
			Correlation: Salt	
Linguistic Distance	-0.432***	-0.483***	-0.371***	-0.206***
	(0.068)	(0.081)	(0.071)	(0.073)
N	18,821	18,821	18,821	18,821
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A4. Remove outliers by linguistic distance

	(1)	(2)	(3)	(4)
			Correlation: Salt	
Linguistic Distance	-0.462***	-0.525***	-0.410***	-0.226***
	(0.077)	(0.080)	(0.074)	(0.075)
N	18,800	18,800	18,800	18,800
			Correlation: Rice	
Linguistic Distance	-0.077***	-0.070***	-0.059***	-0.036***
	(0.014)	(0.010)	(0.015)	(0.009)
N	19,027	19,027	19,027	19,027
			Correlation: Rice	
Linguistic Distance	-0.072***	-0.077***	-0.055***	-0.036***
	(0.019)	(0.011)	(0.019)	(0.010)
N	19,015	19,015	19,015	19,015
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A5. Prices before 1891

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-0.236***	-0.264***	-0.090	-0.032
	(0.049)	(0.043)	(0.068)	(0.051)
N	15,165	15,165	15,165	15,165
			Correlation: Wheat	
Linguistic Distance	-0.081***	-0.148***	-0.058***	-0.039**
	(0.015)	(0.025)	(0.014)	(0.020)
N	13,690	13,690	13,690	13,690
			Correlation: Salt	
Linguistic Distance	-0.490***	-0.672***	-0.392***	-0.261***
	(0.081)	(0.090)	(0.080)	(0.082)
N	19,701	19,701	19,701	19,701
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A6. Prices after 1891

	(1)	(2)	(3)	(4)
			Correlation: Salt	
Linguistic Distance	-0.342***	-0.081**	-0.195***	-0.048**
	(0.058)	(0.033)	(0.057)	(0.021)
N	20,485	20,485	20,485	20,485
			Correlation: Rice	
Linguistic Distance	-0.158***	-0.229***	-0.077**	-0.067*
	(0.024)	(0.028)	(0.032)	(0.038)
N	19,697	19,697	19,697	19,697
			Correlation: Rice	
Linguistic Distance	-0.079***	-0.070***	-0.066***	-0.038***
	(0.017)	(0.013)	(0.016)	(0.009)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A7. Log ρ as outcome

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-0.158***	-0.127***	-0.012	-0.046**
	(0.024)	(0.024)	(0.016)	(0.020)
N	15,648	15,648	15,648	15,648
			Correlation: Salt	
Linguistic Distance	-0.419***	-0.493***	-0.376***	-0.187***
	(0.076)	(0.086)	(0.065)	(0.068)
N	20,615	20,615	20,615	20,615
			Correlation: Rice	
Linguistic Distance	-0.046***	-0.041***	-0.031***	-0.020***
	(0.010)	(0.006)	(0.011)	(0.006)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A8. Centiles of ρ as outcome

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-40.570***	-32.937***	-6.507	-2.753
	(3.538)	(4.127)	(4.277)	(3.325)
N	15,652	15,652	15,652	15,652
			Correlation: Salt	
Linguistic Distance	-29.320***	-30.521***	-25.616***	-11.477***
	(3.762)	(4.294)	(3.875)	(3.693)
N	20,683	20,683	20,683	20,683
			Correlation: Rice	
Linguistic Distance	-23.418***	-20.859***	-13.190***	-6.610***
	(3.138)	(1.876)	(3.631)	(1.879)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A9. $\delta = 0.5$

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Linguistic Distance	-0.333***	-0.189***	-0.030	-0.037**
	(0.039)	(0.025)	(0.022)	(0.019)
N	15,652	15,652	15,652	15,652
		Correlation: Salt		
Linguistic Distance	-0.776***	-0.685***	-0.494***	-0.137*
	(0.084)	(0.084)	(0.081)	(0.077)
N	20,683	20,683	20,683	20,683
		Correlation: Rice		
Linguistic Distance	-0.148***	-0.116***	-0.087***	-0.042***
	(0.020)	(0.012)	(0.020)	(0.012)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A10. Measure distance using largest ethnic group

	(1)	(2)	(3)	(4)
		Correlation: Wheat		
Distance by largest language	-0.206***	-0.141***	-0.038*	-0.047**
	(0.035)	(0.027)	(0.020)	(0.020)
N	15,652	15,652	15,652	15,652
		Correlation: Salt		
Distance by largest language	-0.433***	-0.438***	-0.334***	-0.170***
	(0.063)	(0.070)	(0.061)	(0.058)
N	20,683	20,683	20,683	20,683
		Correlation: Rice		
Distance by largest language	-0.064***	-0.055***	-0.045***	-0.023***
	(0.014)	(0.009)	(0.014)	(0.008)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A11. Cluster by largest ethnic group

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**
	(0.041)	(0.038)	(0.035)	(0.031)
N	15,652	15,652	15,652	15,652
			Correlation: Salt	
Linguistic Distance	-0.494***	-0.555***	-0.422***	-0.224***
	(0.122)	(0.140)	(0.105)	(0.086)
N	20,683	20,683	20,683	20,683
			Correlation: Rice	
Linguistic Distance	-0.083***	-0.073***	-0.056***	-0.035**
	(0.032)	(0.018)	(0.019)	(0.016)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by largest ethnic groups in market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. . Fixed effects are for market i and j.

TABLE A12. Cluster by province

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-0.257***	-0.210***	-0.023*	-0.067**
	(0.046)	(0.043)	(0.012)	(0.032)
N	15,652	15,652	15,652	15,652
			Correlation: Salt	
Linguistic Distance	-0.494***	-0.555***	-0.422***	-0.224*
	(0.176)	(0.175)	(0.134)	(0.126)
N	20,683	20,683	20,683	20,683
			Correlation: Rice	
Linguistic Distance	-0.083**	-0.073***	-0.056***	-0.035*
	(0.038)	(0.023)	(0.016)	(0.018)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by provinces of market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A13. Control for mean absolute log difference

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-0.084***	-0.112***	-0.004	-0.047*
	(0.032)	(0.030)	(0.025)	(0.027)
N	15,652	15,652	15,652	15,652
			Correlation: Salt	
Linguistic Distance	-0.335***	-0.333***	-0.301***	-0.114*
	(0.070)	(0.074)	(0.069)	(0.063)
N	20,683	20,683	20,683	20,683
			Correlation: Rice	
Linguistic Distance	-0.086***	-0.057***	-0.072***	-0.036***
	(0.015)	(0.008)	(0.016)	(0.010)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.

TABLE A14. Control for cost distance

	(1)	(2)	(3)	(4)
			Correlation: Wheat	
Linguistic Distance	-0.257***	-0.210***	-0.023	-0.067**
	(0.035)	(0.036)	(0.025)	(0.031)
N	15,652	15,652	15,652	15,652
			Correlation: Salt	
Linguistic Distance	-0.494***	-0.555***	-0.421***	-0.232***
	(0.074)	(0.085)	(0.072)	(0.075)
N	20,683	20,683	20,683	20,683
			Correlation: Rice	
Linguistic Distance	-0.083***	-0.073***	-0.055***	-0.032***
	(0.017)	(0.010)	(0.018)	(0.010)
N	20,909	20,909	20,909	20,909
Fixed Effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

Notes: ***Significant at 1%, **Significant at 5%, *Significant at 10%. Standard errors clustered by market i and market j in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations, $\ln(distance)$ in km, same province, both coastal, connected to river, rainfall correlation, temperature correlation, and absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, slope, religion, and suitabilities for banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for market i and j.