Linguistic Distance and Market Integration in India

James Fenske and Namrata Kala

The role of cultural distance in market integration, particularly in the developing world, has received relatively little attention. Using prices from more than 200 South Asian markets spanning 1861 to 1921, we show that linguistic distance correlates negatively with market integration. 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, 0.181 for salt, and 0.088 for rice. While factors like genetic distance, literacy gaps, and railway connections are correlated with linguistic distance, they do not fully explain the correlation between linguistic distance and market integration.

Economic historians use market integration as a key measure of economic development (Shiue and Keller 2007; Studer 2008). Although language barriers have been stressed in the macroeconomic literature as inhibiting trade and the diffusion of technology (Spolaore and Wacziarg 2009; Guiso, Sapienza, and Zingales 2009), the role of these variables in market integration within countries, particularly in the developing world, has received comparatively little attention, despite the sizable economic impacts that these barriers can have in other contexts (Spolaore and Wacziarg 2018; Ashraf and Galor 2013). In this article, we consider the economy of colonial India, in which 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?

We collect data from *Wages and Prices in India* on grain and salt prices for 206 South Asian markets between 1861 and 1921. These markets span the territories of modern-day Bangladesh, Burma, India, and

The Journal of Economic History, Vol. 81, No. 1 (March 2021). doi: 10.1017/S0022050720000650

James Fenske is Professor, University of Warwick – Economics, Gibbet Hill Road Coventry CV4 7AL, United Kingdom of Great Britain and Northern Ireland. E-mail: J.Fenske@warwick.ac.uk. Namrata Kala is Assistant Professor, MIT Sloan School of Management, Massachusetts Institute of Technology Cambridge, MA. E-mail: kala@mit.edu.

We are grateful to Latika Chaudhary, Martin Fiszbein, Marc Klemp, Alan Taylor, Romain Wacziarg, and to audiences at the Association for the Study of Religion, Economics, and Culture, George Mason University, Pontificia Universidad Católica de Chile, the University of Manchester, the University of Toulouse, 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.

Pakistan. We merge these markets to populations by language collected from the 1901 colonial census of India. We map these languages into 257 ISO language codes from *Ethnologue*, which also provides us with language trees. 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, 0.189 less in the case of salt, and 0.035 less in the case of rice, relative to means of 0.81 (wheat), 0.54 (salt), and 0.81 (rice) across all market pairs in the data. These are large relative to the coefficients we estimate for physical distance, and suggest a possible role for cultural distance in raising trade costs, even for relatively low-value, homogenous goods.

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. Linguistic distances need not matter exclusively for market integration through language; that is, language itself is one of many imperfect measures of broader ancestral distance. This concept may include shared history, institutions, culture, and norms, among other characteristics (Spolaore and Wacziarg 2016). Language barriers may represent more general barriers to the transmission of vertical traits (Spolaore and Wacziarg 2009, 2018). They may capture differences in tastes, and hence the presence or absence of certain 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 2020; Laval, Patin, and Rueda 2016; Rauch and Trindade 2002). They may work through costs of language or education acquisition (Isphording and Otten 2014; Jain 2017; Laitin and Ramachandran 2016; Shastry 2012). They may correlate with common preferences for public goods, redistribution, and infrastructure (Desmet, Gomes, and Ortuño-Ortín 2020; Desmet, Ortuño-Ortín, and Wacziarg 2012, 2017).

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. Historical differences in literacy across market pairs do correlate with linguistic distance, but do not fully account for its correlation with price integration. Although more linguistically similar market pairs evidence longer periods of time connected to the colonial railway system, this fails to explain away the correlation. Thus, 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, not one of these mechanisms can fully account for the barriers of linguistic cleavages.

Our article 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 similarity predicts greater trade between countries (Melitz and Toubal 2014; Hutchinson 2005; Egger and Lassmann 2012; Anderson and Van Wincoop 2004). More generally, linguistic, religious, and cultural distances across societies correlate with ancestral distance and predict a wide range of economic outcomes (Spolaore and Wacziarg 2018). Within Indian economic history, social divisions of language, caste, and religion have been particularly salient. Industrial segregation was driven by information sharing within ethnolinguistic communities (Gupta 2014). Caste and religious divisions, as well as the preferences of caste, ethnic, and religious elites contributed to reduced spending on schooling, which had effects that persisted until the 1970s (Chaudhary 2009; Chaudhary et al. 2012; Chaudhary and Garg 2015).

Second, we contribute to a literature on market integration and trade. Building on works such as Persson (1999) and Shiue and Keller (2007), several contributions in economic history have measured price integration across markets to compare levels of economic development across regions (Studer 2008; O'Rourke and Williamson 2002; Federico 2011).[1] In the study of Indian economic history, Persaud (2019) has shown that price volatility mattered by spurring international migration. More generally, our work is related to a broader literature on the evolution of trade and market integration throughout history (Pascali 2017; Jacks, Meissner, and Novy 2008; Estevadeordal et al. 2003).

[1] Other studies have used historical price series to measure the responsiveness of prices and welfare to variables such as weather shocks and transportation infrastructure (Jia 2014; Waldinger 2014; Andrabi and Kuehlwein 2010).

We also make a substantial data contribution, digitizing both detailed language data from the colonial census and price data spanning a wider set of markets and commodities (68,181 observations) than addressed by the work of Allen (2007), Andrabi and Kuehlwein (2010), or Studer (2008).

The most similar studies to ours, Falck et al. (2012) and Lameli et al. (2015), use dialect similarity within Germany to predict intra-regional trade and migration. Our work differs from these in several respects. Notably, the linguistic cleavages existing in India are greater than those among the often mutually-intelligible dialects of German. We consider possible roles of genetic distance[2] and transport investment. Finally, we provide evidence from a large and multilingual developing country, cover a longer time period, examine price integration as an outcome, and use a more spatially disaggregated unit of analysis.

HISTORICAL BACKGROUND

Language in South Asia

There are four language families prominently represented in South Asia: Indo-European, Dravidian, Sino-Tibetan, and Austro-Asiatic (Asher 2008). Prior to the arrival of Indo-European languages roughly 3,500 years ago, the sub-continent was predominantly Dravidian-speaking (Asher 2008).

Almost half the world's population speaks an Indo-European language descended from the protolanguage that originated at least 6,000 years ago in eastern Anatolia (Gamkrelidze and Ivanov 1990). These spread throughout Europe and South Asia through both population movement and replacement of languages used by existing populations (Renfrew 1989; Haak et al. 2015). Most speakers of Indo-European languages in South Asia speak Indo-Aryan languages such as Hindi and Bengali. Indo-Aryan languages date back at least as far as 100 BCE (Asher 2008; Emeneau 1956). The principal Dravidian languages became separated no later than 1000 CE, the main literary languages being Telugu, Kannada, Tamil, and Malayalam (Asher 2008). Tamil cave inscriptions date to the second century BC, Malayalam inscriptions to the ninth century AD, Kannada inscriptions to 450 AD, and Telugu places names to the second century AD (Krishnamurti 2003). Austro-Asiatic languages, divided primarily into the Mon-Khmer and Munda branches, predate

[2] See Giuliano, Spilimbergo, and Tonon (2014) as an example for trade among countries.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

the Indo-European languages in South Asia, and may have been present as long as the Dravidian languages (Asher 2008). The small number of Sino-Tibetan speakers in South Asia speak primarily Tibeto-Burman languages (Asher 2008).

Within India, the presence of multiple languages has been shaped by population movements and divergence of relatively isolated speakers (Asher 2008). The rapid adoption of Indo-European languages suggests these had been adopted by the broader Dravidian speaking community as a lingua franca (Krishnamurti 2003), although the Dravidian boundary has been shifting southwards for a very long time, and Dravidian languages were largely absent from the Gangetic valley by 0 AD (Emeneau 1956). Languages in close proximity to each other have influenced each other (Montaut 2005, p. 91). Malayalam uses several Sanskrit words, inflected words, and phrases (Krishnamurti 2003). Indian languages borrow from each other through extensive bilingualism, and Indo-European and Dravidian languages have had grammatical impacts on each other (Krishnamurti 2003; Emeneau 1956). A particular feature of India is the durability of migrant languages, for example, the continued use of Gujurati by communities that have lived in Tamil Nadu for several centuries (Montaut 2005, p. 94).

Markets in Colonial India

The secondary literature on Indian history provides some information on how local prices of foodgrains were determined. Andrabi and Kuehlwein (2010) cite figures demonstrating that production was regionally concentrated, and that most food grains were largely consumed within India. For example, in 1919, the Punjab and the United Provinces accounted for 70 percent of the acreage devoted to growing wheat, while Bengal, Bihar, Orissa, and Madras accounted for 70 percent of the acreage devoted to growing rice. Only 5 percent of wheat and 7 percent of rice was exported beyond India in 1895. Exchange even within India was limited. The non-monetary sector of the economy was large (Kumar 1983), even in 1950 (Chandavarkar 1983).

At the start of our period, 1861, trade costs were high. Land transport was expensive and slow, with food grains largely hauled by oxen walking along dilapidated roads and carrying loads on their backs or in carts (Bhattacharya 1983). In Western India, for example, where few roads existed, trade relied on donkeys, camels, and bullocks (Divekar 1983). Intraregional trade in low-value commodities was possible along rivers, but access to this trade was spatially limited (Derbyshire 1987).

Bullocks required a year to travel the distance that a railway would later cover in a week (McAlpin 1974). Where a lack of roads made wheeled transportation difficult, caravans carried cotton and grain (Roy 2012). Large-scale, long-distance shipments of grain were generally unprofitable (Hurd 1975). The costs of overland transport limited market integration (Kessinger 1983). Migration rates were low and wage convergence among districts over the nineteenth century was slow (Collins 1999). Speed, cost, and seasonality constrained the geographical scope of the commercial orbit of the United Provinces (Derbyshire 1987).

These costs fell during the 60-year time period of our analysis. The telegraph network spread through India in the 1850s and 1870s (Collins 1999). Increasing commercialization benefitted from 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), increased incomes (Donaldson 2018), and reduced famines (Burgess and Donaldson 2010); they are likely to have also increased price co-movement across districts. 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 that lacked railroads to neighboring districts with rail connections.

How did markets themselves work? Bhattacharya (1983) describes prototypical local market places in Eastern India in which farmers sold directly to consumers and middlemen in small quantities, and itinerant traders made small profits exploiting price differences within limited areas. Large farmers served as links among 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. Larger towns gave rise to a stratified system of 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. Company agents contracted with farmers in the villages, while landlords and others lent money to these

farmers and were repaid in grain that they also sold to the commission and buyers' agents. In more remote areas, itinerant traders, including peasants, brought crops to bazaars. At this time, forward trade seldom occurred. Europeans were largely absent from this trade, particularly from local transactions, although they were occasionally company agents and commission agents in railway towns. This helps explain why Europeans, sharing a common language, did not do more to drive market integration and may help explain our results.

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 WWI (McAlpin 1983).

Language in Markets in Colonial India

The languages used in trade varied from market to market, depending on which trading castes were dominant in each location. These are often described in the *Imperial Gazetteer*s for each province.[3] In the Punjab, for example, the multilingual Banias, Khatris, and Aroras who spoke local languages such as Punjabi and Gujarati were dominant in different parts of the province. Predominantly Urdu-speaking Shaikhs and largely Gujarati-speaking Khojas were also important (p. 49). In wheat markets, cultivators themselves traded directly with exporters (p. 87). In Bengal, much of the trade was in the hands of Marwari Agarwals and Oswals, who might often speak local languages. Hindi-speaking Rauniars and Kalwars were more prominent in Bihar (p. 91). In Madras, the Tamil-speaking Chettis and Telugu-speaking Komatis controlled trade in the districts where these languages dominated. Traders themselves were, however, often multilingual, and changed the language used depending on the market. As Montaut (2005, p. 94), drawing on Pandit (1977), puts it:

> The classic example is of the Gujarati merchant one century ago, who uses Kacchi (a dialect of Gujarati) in the local market, Marathi for wider transactions in the region, standard Gujarati for readings, Hindustani when he travels (railway station), Urdu in the mosque, with some Persian and Arabic, but also *sant bhasha* in devotional songs, his variety of Gujarati for family interaction, English when dealing with officials.

[3] *Imperial Gazetteer* of India, Provincial Series, Vol 1. Bengal (1909), Madras (1908), and Punjab (1908). Superintendent of Government Printing.

EMPIRICAL STRATEGY AND DATA

Empirical Strategy

In this article, 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:

$$\rho_{ij}^{p} = \beta^{p} LinguisticDistance_{ij} + x_{ij}^{p\prime}\gamma^{p} + \delta_{i}^{p} + \eta_{j}^{p} + \varepsilon_{ij}^{p}. \quad (1)$$

In Equation (1), ρ_{ij}^{p} is the correlation coefficient for the price of p between markets i and j. $LinguisticDistance_{ij}$, described later, 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, as well as controls for *proximity* (log distance in kilometers between the markets, whether both markets are coastal, and whether both markets are connected by the same river), *geographic similarity* (the correlations in precipitation and temperature between the markets, and their absolute differences in: altitude, latitude, longitude, rainfall, temperature, land quality, ruggedness, malaria, humidity, precipitation, and terrain slope), *agricultural similarity* (absolute differences in suitabilities for growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, or tomato), *other measures of similarity* (whether the markets are in the same province, and their religious distance), and *characteristics of the data* (first year, last year, and the number of years in which the price is available for both markets).

One limitation of our empirical strategy is the possibility that our control variables are measured with greater error than our principal right-hand-side variable of interest, that is, $LinguisticDistance_{ij}$. This could lead to our estimates of β^{p} being overstated. We note, then, that linguistic distance may be interpreted more broadly, for example, as a measure of greater ancestral distance. 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, Gelbach, and Miller 2011). Because of the possible spatial dependence induced by

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

forming every pairwise combination of markets, we show results in the Online Appendix in which we cluster at alternative levels and compute Conley (1999) standard errors.

Data

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

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 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 more than a dozen crops: Arhar Dal, Bajra, Barley, Gram, Jawar, Kangni, Maize, Marua, Rice, Salt, Wheat, Bulrush Millet and Similar, Great Millet and Similar, and Lesser Millets. The data covers both British India and the Princely States. These do not represent all markets in India—almost every populated place would have a market of some sort. Rather, these are markets in which the colonial government collected price data. More populous districts and districts in British India are more likely to appear in the data, and, in provinces such as Coorg that have few districts, at least one district is likely to be present.

In most of our results, we focus on the three most commonly reported prices: rice, wheat, and salt. The data do not allow us to consider differences between different varieties of wheat or salt. However, we also show that estimates of Equation (1) with several other crops produce similar results. 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 uniform fortnightly returns of retail prices were used.[4] So long as there are at least three years in which a price is reported in both markets i and j, 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

[4] We show that results are similar when we use only the period after 1891 (the midpoint of the price data) to compute our dependent variable. We are not worried, then, that differences in how data were collected before and after 1872 drive our results.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

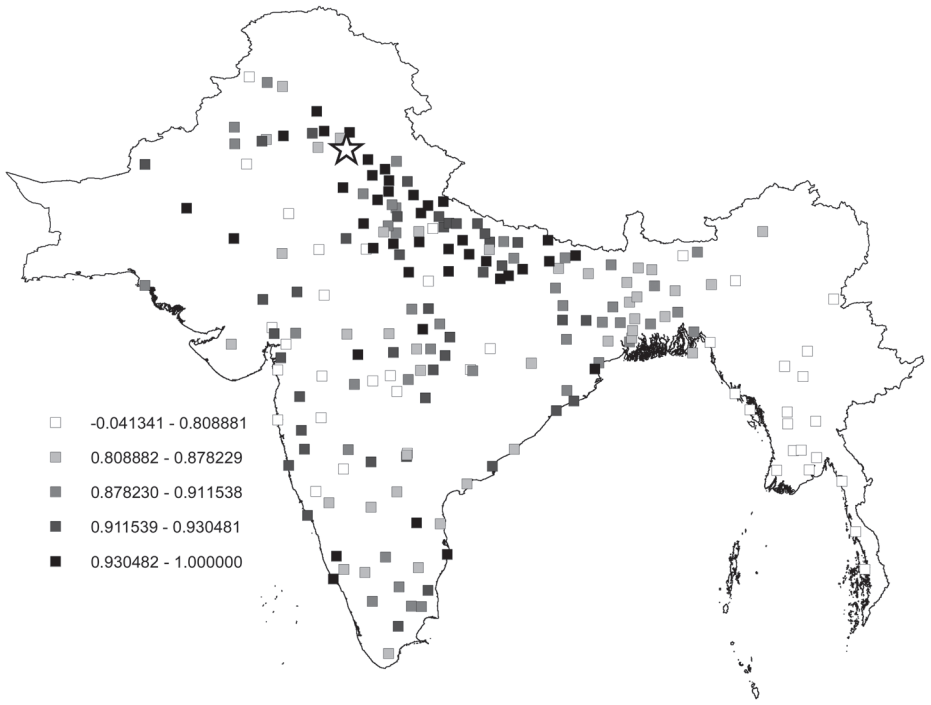

FIGURE 1
LUDHIANA: RICE PRICE CORRELATIONS

Source: *Wages and Prices in India*.

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. Thus, our analysis relies on estimation of Equation (1) to demonstrate that the correlation between linguistic distance and price integration cannot be explained away by other observable differences in proximity or geography.

LINGUISTIC DISTANCE

To compute linguistic distances among 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

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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 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*. The data do not, unfortunately, mention second languages.

To compute the distances among these languages, we turn to *Ethnologue*. Every language in this source is categorized using a language tree with a maximum number of 15 branches. These classifications are based on several sources, the most important of which is Frawley (2003). Such "cladistic" measures have become widely used in economics (Desmet, Ortuño-Ortín, and Wacziarg 2012; Gomes 2014).[5]

Following Esteban, Mayoral, and Ray (2012), we take the distance d_{mn} between any two languages m and n as:

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

Similarly following Esteban, Mayoral, and Ray (2012), we choose $\delta = 0.05$ as a baseline and use $\delta = 0.5$ for robustness. To aggregate these to distances among markets, given population shares of languages m and n in each district i and j of s_{mi} and s_{nj}, we follow Spolaore and Wacziarg (2009) and compute linguistic distance among districts as:

$$LD_{ij} = \sum_m \sum_n (s_{mi} \times s_{nj} \times d_{mn}). \tag{3}$$

In Figure 2, we map the linguistic distances among 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

[5] Although alternative distance measures exist based on phonetic similarity of languages (Dickens 2018), these would be measured with considerable error in our data, given the large number of languages in our data for which the phonetic word lists of the Automated Similarity Judgment Program are either missing or incomplete. (We do, however, report results using these as an alternative measure). Under this classification system, for example, Punjabi is coded as Indo-European, Indo-Iranian, Indo-Aryan, Intermediate Divisions, Western, and Panjabi.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

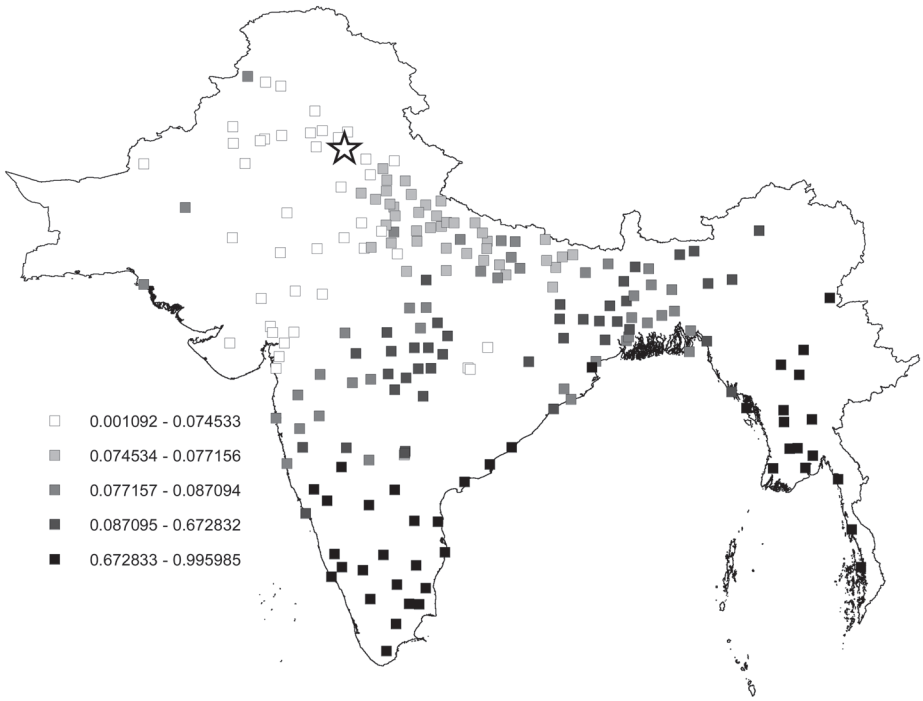

FIGURE 2
LUDHIANA: LINGUISTIC DISTANCES

Source: Census of India 1901.

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 kilometers between Ludhiana and Dacca is 7.40, whereas it is 7.76 between Ludhiana and Madurai.

ADDITIONAL CONTROLS

Some of our control variables are computed directly. Distance in kilometers is computed using the latitude and longitude of the market. "Both coastal" and "both connected by the same river" indicators 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 is based on the provinces that contained each market in 1901. The "religious distance" variable is computed using

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

the same equation as Equation (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_{qr} between any religion q and r is 1 if $q \neq r$ and 0 if $q = r$.[6]

Data on land quality are taken from Ramankutty et al. (2002) and have been used in several economic studies, such as Michalopoulos (2012) and Ashraf and Galor (2011).[7] 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).[8] Our measure of "malaria prevalence" was originally created by Kiszewski et al. (2004).[9] Altitude data are taken from the Consultative Group for International Agricultural Research's Shuttle Radar Topography Mission 30 dataset.[10] Means of precipitation, temperature, and suitabilities for specific crops are taken from the Food and Agriculture Organization's Global Agro-Ecological Zones data portal.[11] Similar suitability measures have been used by Alesina, Giuliano, and Nunn (2013) and Alsan (2015). Correlations in rainfall are computed using the Matsuura and Willmott (2007) gridded series.[12] 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.[13]

Like many studies that control for geographic confounders with historical outcome variables, we are compelled to use present-day raster data (e.g., Alsan (2015) and Nunn and Puga (2012)). We expect that this will add measurement error to our right-hand-side variables, but that it is unlikely this measurement error will induce spurious correlation between linguistic distance and market integration. For the variables that require geographic data (i.e., the coastal and river indicators, as well as those using raster data), we begin with a district map for modern India.[14] We

[6] If, as an alternative, we collapse Islam, Judaism, and Christianity into a single category, results are numerically indistinguishable because of the negligible share of Jews and Christians in the population. We omit these results for space.

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

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

[9] We are grateful to Marcella Alsan for providing us with these data.

[10] http://www.diva-gis.org/gdata

[11] http://www.fao.org/nr/gaez/en/

[12] http://climate.geog.udel.edu/~climate

[13] https://crudata.uea.ac.uk/cru/data/hrg/tmc/grid_10min_reh.dat.gz

[14] In particular, we use the boundaries reported by www.gadm.org.

TABLE 1
SUMMARY STATISTICS

	(1) Mean	(2) Standard Deviation	(3) Min.	(4) Max.	(5) N
Correlation: Wheat	0.81	0.22	–1	1	15,652
Correlation: Salt	0.54	0.41	–0.78	1	20,909
Correlation: Rice	0.81	0.16	–0.25	1	20,909
Linguistic Distance (d=0.05)	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

Source: See the text.

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 have a unique match between the market and the modern district polygon. Otherwise, we match all districts that split from the erstwhile district that previously shared the name of the market to that market.

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; this limits arbitrage opportunities in response to shocks, causing lower 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).

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

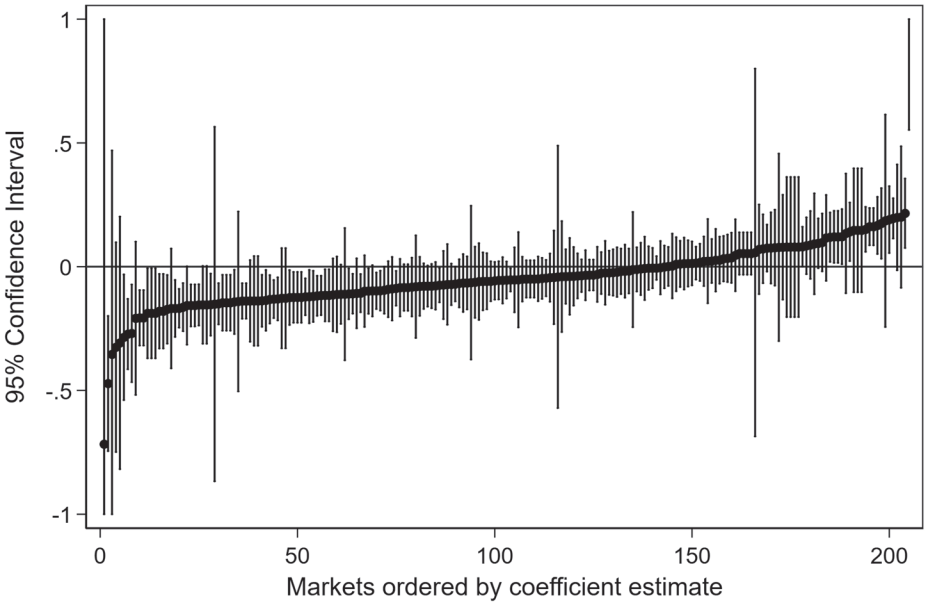

FIGURE 3
RESULTS BY MARKET: WHEAT

Source: Authors' estimates of Equation (4).

RESULTS

Results by Market

Before presenting estimates of Equation (1), we present preliminary descriptive evidence.[15] For each market *i* in our data, we estimate:

$$\rho_{ij}^{p} = \beta_{i}^{p} LinguisticDistance_{ij} + x_{ij}^{p\prime}\gamma^{p} + \varepsilon_{ij}^{p}. \quad (4)$$

In Equation (4), ρ_{ij}^{p} and x_{ij}^{p} are defined as in Equation (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 percent confidence intervals in Figures 3, 4, and 5. For each of the three major crops, the majority of coefficients

[15] Fenske and Kala (2020) provide data and code to replicate all analyses in this paper.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

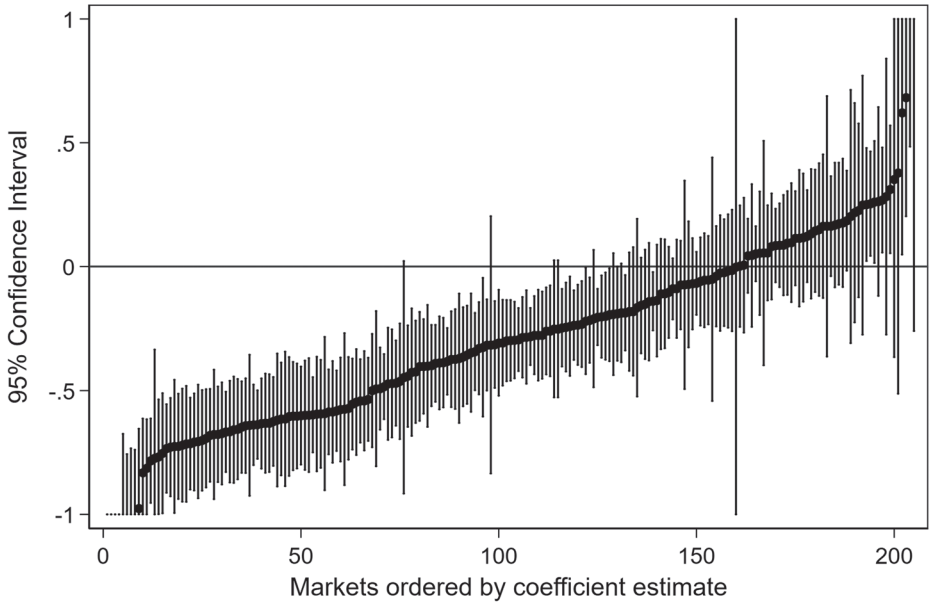

FIGURE 4
RESULTS BY MARKET: SALT

Source: Authors' estimates of Equation (4).

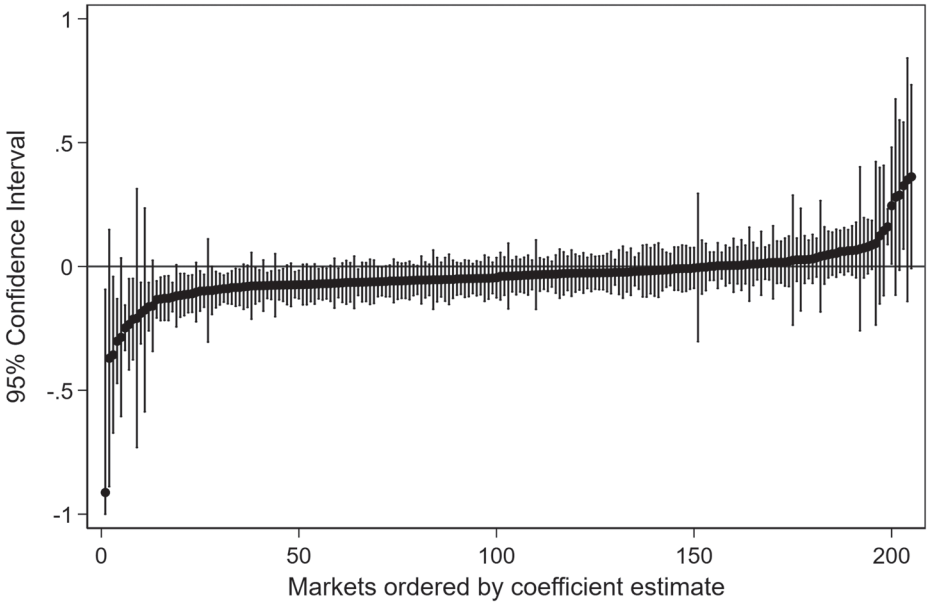

FIGURE 5
RESULTS BY MARKET: RICE

Source: Authors' estimates of Equation (4).

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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, Equation (1) yields estimates of β^p that capture a central tendency in the sample.

Main Results

In Table 2, we present our main estimates of Equation (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, 0.181 standard deviations for salt, and 0.088 standard deviations for rice.

It is striking that the coefficients and standardized magnitudes are largest for salt. Not only are salt markets less integrated in the data, in that they have lower mean correlation coefficients, there is also more dispersion in integration for salt, in that the standard deviation of the correlation coefficients across market pairs is larger. Salt was a differentiated good that could only be produced in a small number of locations (Donaldson 2018). Further, in order to facilitate the taxation of salt, the British constructed an Inland Customs Line, which incorporated the Great Hedge of India, in order to prevent salt smuggling (Moxham 2001).

An alternative approach to magnitudes is to divide $\widehat{\beta^p}$ 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 percent for wheat, 1,328 percent for salt, and 210 percent for rice. At the mean distance across pairs within our sample (1,154 kilometers), this would correspond to distance increases of 9,101, 15,326, and 2,418 kilometers, respectively, all of which would be out of sample. These large numbers are driven in part by the small coefficients estimated on distance once additional controls are included.

In Online Appendix Table A4, we compare the pairwise correlations between our outcome variables and the measures of physical and linguistic distance. Both distance measures enter significantly and negatively on their own and, if both are put on the right-hand side at once, both continue to enter negatively and significantly, while the coefficient

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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)
Observations	15,652	15,652	15,652	15,652
R-squared	0.139	0.762	0.580	0.806
			Correlation: Salt	
Linguistic distance	–0.484***	–0.392***	–0.384***	–0.189***
	(0.061)	(0.072)	(0.051)	(0.044)
Observations	20,909	20,909	20,909	20,909
R-squared	0.216	0.708	0.566	0.791
			Correlation: Rice	
Linguistic distance	–0.083***	–0.073***	–0.056***	–0.035***
	(0.017)	(0.010)	(0.018)	(0.010)
Observations	20,909	20,909	20,909	20,909
R-squared	0.045	0.834	0.282	0.868
Fixed effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: 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, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for markets *i* and *j*.
Source: See the text.

on each is reduced slightly. Both have similar R-squared values when included as right-hand-side variables alone, and including both on the right-hand side increases the R-squared.

MECHANISMS

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

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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 imparted across generations in driving dissimilarities in economic outcomes across populations (Spolaore and Wacziarg 2009, 2018). Alternatively, differences in language may proxy for differences in tastes, which, in turn, shape prices and the volume of trade (Atkin, 2013, 2016). Where these taste-based differences lead to a thin local market for a given good, we might anticipate prices that do not track those in other South Asian markets. Similarly, if there are fixed costs of arbitrage between two markets, the limited size of the market for an unpopular product will reduce the returns to arbitrage.

Another branch of the economic literature suggests 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; Allen 2014). Alternatively, linguistic distance may act as a barrier to flows of people, who are likely to be put off by migration costs, the difficulty of establishing business connections, or by xenophobia (Bai and Kung 2020; Falck et al. 2012; Lameli et al. 2015; Rauch and Trindade 2002; Iwanowsky 2018). These mechanisms would lead to missing or costly links in the network connecting any two markets.

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, pp. 188–96) mentions several such "middlemen minorities."[16] These groups, Divekar (1983) argues, contributed to the

[16] His list includes the Marwaris, Gujaratis, Parsis, Sindhis, Chettiars, Khatris, Aroras, Multanis, Bhatias, Khojas, Lohanas, Bohras, Memons, Banias, Pathans, Vanis, Shravaks, Agarwals, Maheshwaris, Oswals, Khandelwals, and Porwals. Roy (2014) similarly discusses the role of Marwaris, Banias, Parsis, and Khojas. Divekar (1983) adds to this the Afghans, Voras, Lingayat Banjigs, Komtis, and Vanjaris. Kumar (1983) and McAlpin (1974), similarly, highlight the role of the Banjaras.

"unification of markets in India." They adopted new forms of business partnership and circulated information over wide regions. If the costs of one group maintaining a presence in a given market due to its linguistic dissimilarity are greater, this would be expected to increase transactions costs with other markets in which they are present.

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 (2017), 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 (Desmet, Gomes, and Ortuño-Ortín 2020; Desmet, Ortuño-Ortín, and Wacziarg 2012, 2017). If these public goods and infrastructure investments affect trade costs, they may help explain our main result.

Mechanisms: Evidence

GENETIC DISTANCE

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 among the markets in our data. We show that, while linguistic distance and genetic distance are correlated, neither one is a "sufficient statistic" that fully accounts for the coefficient of the other.

We obtain data on genetic distance from Pemberton, DeGiorgio, and Rosenberg (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 5,795 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, Miso, 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.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

Following Spolaore and Wacziarg (2009), given population shares of groups m and n in districts i and j of s_{mi} and s_{nj} with genetic distance F_{ST}^{mn}, we compute genetic distance among districts as:

$$GD_{ij} = \sum_m \sum_n (s_{mi} \times s_{nj} \times F_{ST}^{mn}). \quad (5)$$

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 Online Appendix Figure A1. This has many similarities to Figure 2. Other regions of South Asia that are proximate to the Punjab are more genetically similar, although 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 Equation (1), first with genetic distance as an outcome, and second with genetic distance as an additional control. We report the results in Table 3. Linguistic and genetic distance are correlated, even conditional on our baseline fixed effects and controls.[17] 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. With fixed effects and controls, the change in coefficient on linguistic distance is slight when compared with Table 2. These results imply that, while linguistic distance may indeed proxy for other differences across populations, its relationship with market integration cannot be fully accounted for by the additional transaction costs imposed by barriers to the diffusion of beliefs, traditions, and practices stemming from ancestral distance.

COARSE AND FINE DISTINCTIONS

We show that it is the highest-level distinctions in our data, such as those between Indo-European and Dravidian languages, that drive our results. This is, however, a crude proxy, and we cannot rule out the possibility that languages here proxy for past patterns of migration and state formation that themselves shaped markets and trade routes.

[17] In the sample of pairwise comparisons among the 24 ethnic groups in Pemberton, DeGiorgio, and Rosenberg (2013), avoiding duplicates and self-comparisons by keeping only ij pairs where $i < j$, the correlation between genetic and linguistic distance is positive but small, with $\rho = 0.1216$.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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)
Observations	21,115	21,115	21,115	21,115
R-squared	0.012	0.857	0.360	0.895
		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)
Observations	15,652	15,652	15,652	15,652
R-squared	0.142	0.769	0.580	0.806
		Correlation: Salt		
Linguistic distance	–0.465***	–0.367***	–0.371***	–0.194***
	(0.064)	(0.079)	(0.052)	(0.043)
Genetic distance X 100	–0.415***	–0.234**	–0.287***	0.195**
	(0.126)	(0.096)	(0.100)	(0.081)
Observations	20,909	20,909	20,909	20,909
R-squared	0.242	0.710	0.574	0.792
		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)
Observations	20,909	20,909	20,909	20,909
R-squared	0.074	0.838	0.291	0.869
Fixed effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: 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, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for markets *i* and *j*.
Source: See the text.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

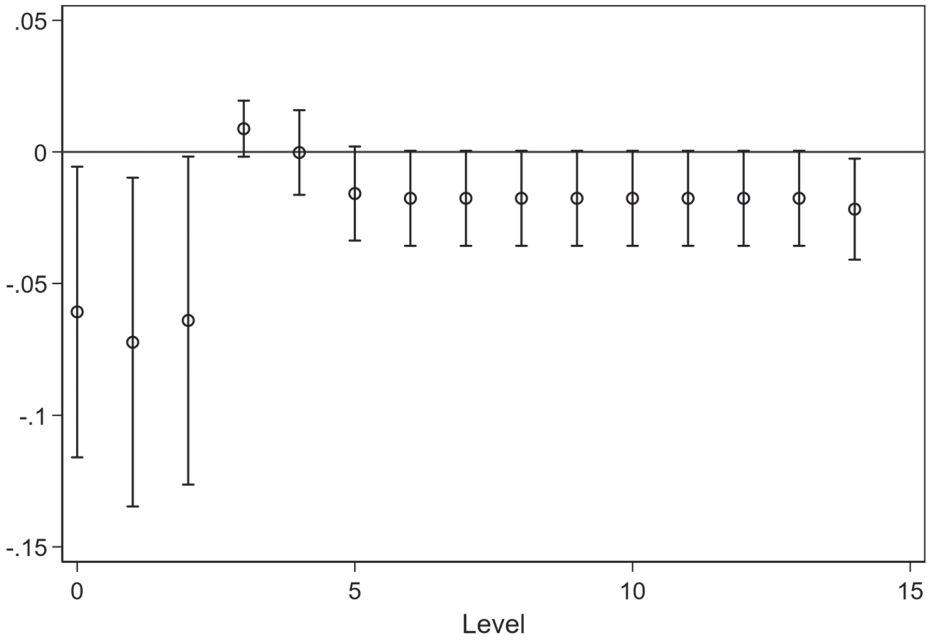

FIGURE 6
RESULTS BY LEVEL: WHEAT

Source: Authors' estimates.

Recall that, in our baseline analyses, we computed the distance between any two languages m and n as:

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

While this follows the convention in the literature, it does not allow us to distinguish whether coarser distinctions (e.g., those between Indo-European and Dravidian languages) or lesser divisions (e.g, those between Bengali and Punjabi) drive our results. We replace d_{mn} with a dummy for having $\leq N$ shared branches, for $N = \{1, ..., 15\}$. We re-estimate Equation (1), and present our results in Figures 6, 7, and 8. 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. Indeed, we show in Online Appendix Table A5 that limiting our sample only to district pairs in which the dominant language in both districts is Indo-European leads to coefficient estimates on linguistic distance that, while still negative, are generally insignificant and less robust across specifications. That is, our results are driven by coarser language distinctions, particularly those that separate major language families.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

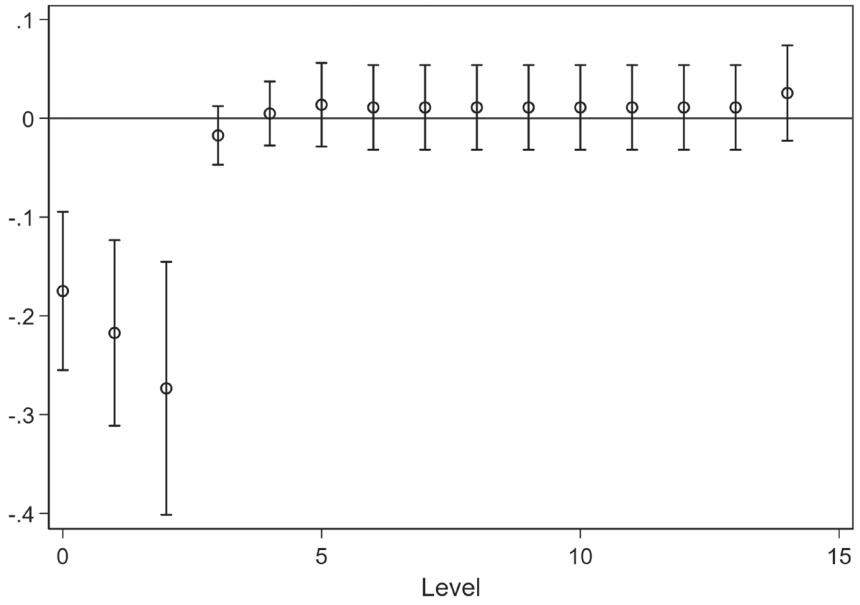

FIGURE 7
RESULTS BY LEVEL: SALT

Source: Authors' estimates.

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 than finer distinctions. These results suggest that our main results derive from divisions on the scale of Gujarati-Tamil, Gujarati-Yiddish, and Gujarati-Balochi, rather than from finer distinctions as those among Gujarati and Hindi, Nepali, or Jandavra. These coarser distinctions are those that have been shown before to correlate with conflict, redistribution, and public goods provision—suggesting they are correlated with deeper differences in preferences—as opposed to finer distinctions that inhibit coordination and integration (Desmet, Ortuño-Ortín, and Wacziarg

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

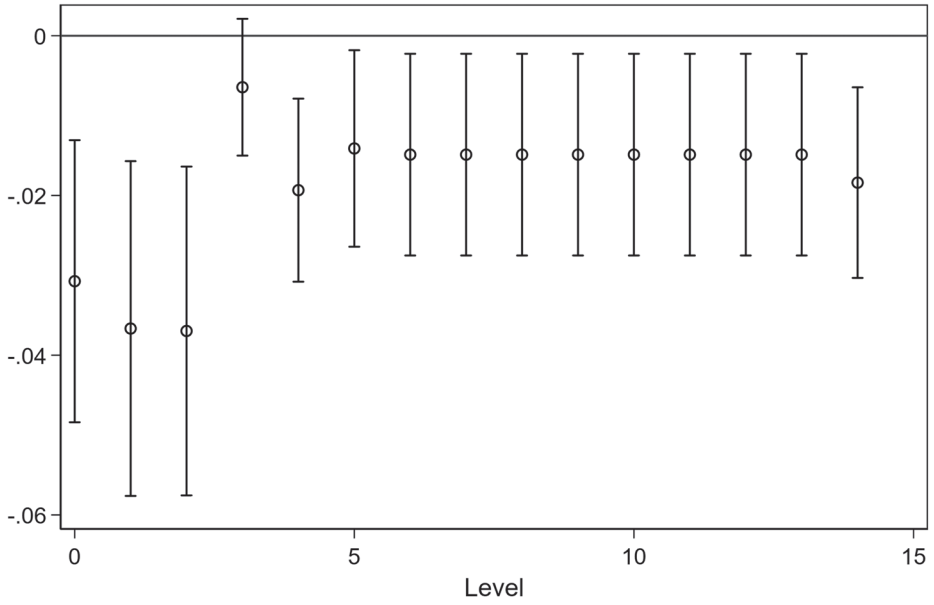

FIGURE 8
RESULTS BY LEVEL: RICE

Source: Authors' estimates.

2012). This is suggestive evidence that our results are driven not simply by ease of communication, but also by more fundamental differences in preferences.

MISSING MARKETS

To test whether missing markets, due, for example, to differences in tastes drive the correlation between linguistic distance and market integration, we evaluate whether linguistic distance predicts whether two given markets report a certain good's price in the same year, and whether markets that are more linguistically distant from their neighbors experience more volatile prices. When we look at the situation for major crops, we find little evidence of missing markets increasing with linguistic distance. Only limited evidence suggests that prices are more variable at markets that are more linguistically different from those around them.

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 Equation (1), except that we now take N_{ij}^{p} as the dependent variable, and no longer control for minimum year,

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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)
Observations	21,115	21,115	21,115	21,115
R-squared	0.429	0.928	0.562	0.936
		Observations: Salt		
Linguistic distance	–1.304	0.004	–3.279*	–0.017
	(1.278)	(0.071)	(1.868)	(0.157)
Observations	21,115	21,115	21,115	21,115
R-squared	0.003	0.954	0.212	0.954
		Observations: Rice		
Linguistic distance	–1.441	0.011	–3.126	–0.097
	(1.316)	(0.085)	(1.938)	(0.165)
Observations	21,115	21,115	21,115	21,115
R-squared	0.003	0.954	0.205	0.955
Fixed effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: 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, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for markets *i* and *j*.
Source: See the text.

maximum year, or the number of common observations. Results are presented in Table 4. There is only weak evidence of missing markets correlating with linguistic distance; while we find a negative correlation between linguistic distance and N_{ij}^{p} for wheat, no such correlation is available for salt or rice. We find similar failures of linguistic distance to predict N_{ij}^{p} when using lesser crops from the data such as barley and maize, although we do not report these here. One explanation of the different result for wheat is the greater variability of the outcome variable: the standard deviation of the number of common years for wheat is 22.6, versus 8.8 for salt and 9.7 for rice. That is, as wheat is reported less often in many markets, there is more variation to be explained.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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. Our logic here is that linguistic distance from neighbors may lead to more volatile prices because of reduced trade and arbitrage. For each market i, we keep the other markets within 500 kilometers and take the average of their linguistic distance from i (denoted $\overline{LinguisticDistance_{ij}}$) as well as the average of the controls (denoted $\overline{x_{ij}^{p}}$). We estimate:

$$CV_i^p = \beta^p \overline{LinguisticDistance_{ij}} + \overline{x_{ij}^{p}}'\gamma^p + \varepsilon_i^p. \quad (6)$$

In Equation (6), CV_i^p is the coefficient of variation of the price of product p at market i. We estimate Equation (6) by ordinary least squares (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. The differences by crop here are somewhat puzzling, as it is rice prices that are most volatile in our data, as measured by the coefficient of variation.

TRADING COMMUNITIES

To evaluate whether the presence of trading networks sharing a common tongue drives our results (e.g., as might be the case 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.

We focus on one group that has received particular attention in the literature: the Marwaris. By 1920, between 200,000 and 400,000 Marwaris, most of them working as traders, lived outside of the Rajputana Agency (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 Equation (1) with $AD_{ij}^{Marwari}$ as both an outcome and as a control. That is, we test whether linguistic distance predicts the colocation of Marwaris

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

TABLE 5
MISSING MARKETS: VOLATILITY

	(1) CV: Wheat	(2) CV: Salt	(3) CV: Rice
Linguistic distance	0.127***	0.030	–0.113
	(0.049)	(0.049)	(0.279)
Observations	178	205	205
R-squared	0.528	0.400	0.121

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: Robust standard errors in parentheses. All regressions are OLS and include a constant. Controls are minimum year, maximum year, number of observations and averages of ln(distance) in km, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato.
Source: See the text.

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 differences in the presence of this trading community, and little evidence that it explains price integration.

Results are similar if we perform the same exercise for the other communities listed, although we do not report these for space. While we cannot observe all these communities in our data, several are recorded in the census either as linguistic or religious groups. In particular, we are able to observe the Parsis, Afghanis, Gujaratis, Khatris, Memons, Multanis, and Sindhis. We also observe the Vanis, but they are not present in the markets in our data. Since the English could also be potentially thought of as another migrant mercantile community, we also consider their presence. Results are again similar, and again not reported, using the English. Our results are particularly unlikely to be explained by the spread of the English language: less than one-tenth of 1 percent of the population in the 1901 census is recorded as "English" by language.

Alternatively, if we replace the absolute difference in the population share of a minority group with the maximum for a market pair, results are very similar. Because a group is often present in one market and not another, the maximum across a pair is highly correlated with the absolute difference in shares. Similarly, we find little correlation between

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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)
Observations	21,115	21,115	21,115	21,115
R-squared	0.004	0.984	0.263	0.984
		Correlation: Wheat		
Linguistic distance	–0.255***	–0.210***	–0.023	–0.067**
	(0.035)	(0.036)	(0.025)	(0.030)
Absolute difference in Marwaris share	0.066***	0.021*	–0.003	0.030*
	(0.014)	(0.011)	(0.016)	(0.017)
Observations	15,652	15,652	15,652	15,652
R-squared	0.142	0.762	0.580	0.806
		Correlation: Salt		
Linguistic distance	–0.498***	–0.391***	–0.360***	–0.189***
	(0.060)	(0.072)	(0.051)	(0.044)
Absolute difference in Marwaris share	–0.571***	–0.274*	–0.425***	–0.174
	(0.068)	(0.152)	(0.083)	(0.149)
Observations	20,909	20,909	20,909	20,909
R-squared	0.260	0.709	0.584	0.791
		Correlation: Rice		
Linguistic distance	–0.085***	–0.073***	–0.054***	–0.035***
	(0.017)	(0.010)	(0.017)	(0.010)
Absolute difference in Marwaris share	–0.074	–0.040***	–0.028	0.015
	(0.065)	(0.012)	(0.073)	(0.013)
Observations	20,909	20,909	20,909	20,909
R-squared	0.050	0.834	0.283	0.868
Fixed effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: 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, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for markets *i* and *j*.
Source: See the text.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

linguistic distance and the minimum presence of a trading community across a market pair, and our results are not generally sensitive to controlling for this minimum. Again, we omit these results for space.

LITERACY

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, this does not diminish the correlation of linguistic distance with market integration.

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

INFRASTRUCTURE

Finally, we 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, nonetheless, this does not fully account for our main result.

Following a procedure similar to Donaldson (2018), we use the 1934 edition of *History of Indian Railways Constructed and In Progress* to 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 Equation (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; however, this does little to predict price correlations or explain away their correlation with linguistic distance. One possible 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).

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)
Observations	20,503	20,503	20,503	20,503
R-squared	0.193	0.808	0.504	0.837
		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)
Observations	15,125	15,125	15,125	15,125
R-squared	0.139	0.761	0.579	0.805
		Correlation: Salt		
Linguistic distance	–0.339***	–0.323***	–0.327***	–0.164***
	(0.052)	(0.054)	(0.047)	(0.040)
Difference in literacy 1921	–0.016***	–0.012***	–0.008***	–0.006***
	(0.002)	(0.003)	(0.002)	(0.002)
Observations	20,300	20,300	20,300	20,300
R-squared	0.343	0.732	0.589	0.800
		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)
Observations	20,300	20,300	20,300	20,300
R-squared	0.155	0.836	0.347	0.869
Fixed effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: 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, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for markets *i* and *j*.
Source: See the text.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

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)
Observations	21,115	21,115	21,115	21,115
R-squared	0.009	0.850	0.170	0.853
	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)
Observations	15,652	15,652	15,652	15,652
R-squared	0.140	0.762	0.580	0.806
	Correlation: Salt			
Linguistic distance	–0.473***	–0.391***	–0.381***	–0.189***
	(0.060)	(0.072)	(0.051)	(0.044)
Years both connected to railroad	0.002***	0.001***	0.001	0.000
	(0.001)	(0.000)	(0.001)	(0.000)
Observations	20,909	20,909	20,909	20,909
R-squared	0.227	0.709	0.567	0.791
	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)
Observations	20,909	20,909	20,909	20,909
R-squared	0.048	0.834	0.282	0.868
Fixed effects	No	Yes	No	Yes
Controls	No	No	Yes	Yes

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: 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, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for markets *i* and *j*.
Source: See the text.

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

TABLE 9
ALTONJI-ELDER-TABER STATISTICS

	(1)	(2)	(3)
		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	4.245	3.849	0.640
Baseline: ln(distance)	2.289	–9.983	1.205
		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

* = Significant at the 10 percent level.
**= Significant at the 5 percent level.
*** = Significant at the 1 percent level.
Notes: 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, 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 growing banana, chickpea, cocoa, cotton, groundnut, dryland rice, oil palm, onion, soybean, sugar, tea, wetland rice, white potato, wheat, and tomato. Fixed effects are for markets *i* and *j*.
Source: See the text.

ROBUSTNESS

Selection on Unobservables

In this section, we demonstrate the robustness of our results to selection on unobservables. We present a number of additional exercises in the Online Appendix.

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

$$AET = \frac{\beta^{FullControls}}{\beta^{RestrictedControls} - \beta^{FullControls}}. \quad (7)$$

We report results 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

https://doi.org/10.1017/S0022050720000650 Published online by Cambridge University Press

sensitive to controls regardless of what is in the base set of controls, but are not as sensitive to the addition of fixed effects. Results for salt and rice 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. That is, we find that the estimate of β is sensitive to controls for wheat, while for salt and rice, the estimate of β is no longer sensitive to controls once ln(*Distance*) has been included.

CONCLUSION

In this article, we have shown that markets in colonial South Asia that were more linguistically distant from each other displayed less market integration, conditional on many other measures, including distance, literacy gaps, transportation links, and measures of dissimilarity. This finding holds across multiple products and markets, and survives several sensitivity checks. Genetic distance and lack of railway connections may help explain these results, but on their own, these factors do not explain the lack of market integration. There is less evidence for missing markets and presence of trading communities as mechanisms. The results show 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 that enhance or impede market integration is substantial relative to other factors such as physical distance. More linguistically-similar markets are more likely to have been connected earlier via transport infrastructure (the colonial railway system), but this connection alone does not explain away the coefficient. These results indicate the importance and persistence of cultural differences in market integration, trade, and price volatility. Testing whether markets with greater gains from trade learn the languages necessary for trade over time, and whether newer information and communication technologies reduce the importance of linguistic distance, remain important questions for future work.

REFERENCES

Adams, John, and Robert Craig West. "Money, Prices, and Economic Development in India, 1861–1895." *Journal of Economic History* 39, no. 1 (1979): 55–68.

Alesina, Alberto, Paola Giuliano, and Nathan Nunn. "On the Origins of Gender Roles: Women and the Plough." *Quarterly Journal of Economics* 128, no. 2 (2013): 469–530.

Allen, Robert C. "India in the Great Divergence." *The New Comparative Economic History: Essays in Honor of Jeffrey G. Williamson* (2007): 9–32.
Allen, Treb. "Information Frictions in Trade." *Econometrica* 82, no. 6 (2014): 2041–83.
Alsan, Marcella. "The Effect of the TseTse Fly on African Development." *American Economic Review* 105, no. 1 (2015): 382–410.
Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113, no. 1 (2005): 151–84.
Anderson, James E., and Eric Van Wincoop. "Trade Costs." *Journal of Economic Literature* 42, no. 3 (2004): 691–751.
Andrabi, Tahir, and Michael Kuehlwein. "Railways and Price Convergence in British India." *Journal of Economic History* 70, no. 2 (2010): 351–77.
Asher, Ronald E. "Language in Historical Context." *Language in South Asia* (2008): 31–48.
Ashraf, Quamrul, and Oded Galor. "Dynamics and Stagnation in the Malthusian Epoch." *American Economic Review* 101, no. 5 (2011): 2003–41.
———. "Genetic Diversity and the Origins of Cultural Fragmentation." *American Economic Review* 103, no. 3 (2013): 528–33.
Atkin, David. "Trade, Tastes, and Nutrition in India." *American Economic Review* 103, no. 5 (2013): 1629–63.
———. "The Caloric Costs of Culture: Evidence from Indian Migrants." *American Economic Review* 106, no. 4 (2016): 1144–81.
Bai, Ying, and James Kung. "Surname Distance and Technology Diffusion: The Case of the Adoption of Maize in Late Imperial China." Working Paper. Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, 2020.
Bellows, John, and Edward Miguel. "War and Local Collective Action in Sierra Leone." *Journal of Public Economics* 93, no. 11 (2009): 1144–57.
Bhattacharya, S. "Regional Economy (1757–1857): Eastern India." *Cambridge Economic History of India* 2 (1983): 270–95.
Burgess, Robin, and Dave Donaldson. "Can Openness Mitigate the Effects of Weather Shocks? Evidence from India's Famine Era." *American Economic Review* 100, no. 2 (2010): 449–53.
Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. "Robust Inference with Multiway Clustering." *Journal of Business & Economic Statistics* 29, no. 2 (2011): 238–49.
Chandavarkar, Anand G. "Money and Credit, 1858–1947." *Cambridge Economic History of India* 2 (1983): 762–803.
Chaudhary, Latika. "Determinants of Primary Schooling in British India." *Journal of Economic History* 69, no. 1 (2009): 269–302.
Chaudhary, Latika, and Manuj Garg. "Does History Matter? Colonial Education Investments in India." *Economic History Review* 68, no. 3 (2015): 937–61.
Chaudhary, Latika, Aldo Musacchio, Steven Nafziger, and Se Yan. "Big BRICs, Weak Foundations: The Beginning of Public Elementary Education in Brazil, Russia, India, and China." *Explorations in Economic History* 49, no. 2 (2012): 221–40.
Collins, William J. "Labor Mobility, Market Integration, and Wage Convergence in Late 19th Century India." *Explorations in Economic History* 36, no. 3 (1999): 246–77.

Conley, Timothy G. "GMM Estimation with Cross Sectional Dependence." *Journal of Econometrics* 92, no. 1 (1999): 1–45.

Derbyshire, Ian D. "Economic Change and the Railways in North India, 1860–1914." *Modern Asian Studies* 21, no. 3 (1987): 521–45.

Desmet, Klaus, Joseph Flavian Gomes, and Ignacio Ortuño-Ortín. "The Geography of Linguistic Diversity and the Provision of Public Goods." *Journal of Development Economics* 143 (2020): 102384.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg. "The Political Economy of Linguistic Cleavages." *Journal of Development Economics* 97, no. 2 (2012): 322–38.

———. "Culture, Ethnicity and Diversity." *American Economic Review* 107, no. 9 (2017): 2479–2513.

Dickens, Andrew. "Ethnolinguistic Favouritism in African Politics." *American Economic Journal: Applied Economics* 10, no. 3 (2018): 370–402.

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

Donaldson, Dave. "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure." *American Economic Review* 108, nos. 4–5 (2018): 899–934.

Egger, Peter H., and Andrea Lassmann. "The Language Effect in International Trade: A Metaanalysis." *Economics Letters* 116, no. 2 (2012): 221–24.

Emeneau, Murray B. "India as a Lingustic Area." *Language* 32, no. 1 (1956): 3–16.

Esteban, Joan, Laura Mayoral, and Debraj Ray. "Ethnicity and Conflict: An Empirical Study." *American Economic Review* 102, no. 4 (2012): 1310–42.

Estevadeordal, Antoni, Brian Frantz, Alan M. Taylor, et al. "The Rise and Fall of World Trade, 1870–1939." *Quarterly Journal of Economics* 118, no. 2 (2003): 359–407.

Falck, Oliver, Stephan Heblich, Alfred Lameli, and Jens Südekum. "Dialects, Cultural Identity, and Economic Exchange." *Journal of Urban Economics* 72, no. 2 (2012): 225–39.

Federico, Giovanni. "When Did European Markets Integrate?" *European Review of Economic History* 15, no. 1 (2011): 93–126.

Fenske, James, and Namrata Kala. "Replication: Linguistic Distance and Market Integration in India." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-10-12. https://doi.org/10.3886/E124121V1.

Frawley, William J. *International Encyclopedia of Linguistics*. Vol. 4. Oxford University Press, 2003.

Gamkrelidze, Thomas V., and Vyacheslav V. Ivanov. "The Early History of Indo-European Languages." *Scientific American* 262, no. 3 (1990): 110–17.

Giuliano, Paola, Antonio Spilimbergo, and Giovanni Tonon. "Genetic Distance, Transportation Costs, and Trade." *Journal of Economic Geography* 14, no. 1 (2014): 179–98.

Gomes, Joseph Flavian. "The Health Costs of Ethnic Distance: Evidence from Sub-Saharan Africa." ISER Working Paper Series No. 2014-33, Colchester, UK, 2014.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. "Cultural Biases in Economic Exchange?" *Quarterly Journal of Economics* 124, no. 3 (2009): 1095–131.

Gupta, Bishnupriya. "Discrimination or Social Networks? Industrial Investment in Colonial India." *Journal of Economic History* 74, no. 1 (2014): 141–68.

Haak, Wolfgang, et al. "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe." *Nature* 522, no. 7555 (2015): 207–11.

Hurd, John. "Railways and the Expansion of Markets in India, 1861–1921." *Explorations in Economic History* 12, no. 3 (1975): 263–88.

Hutchinson, William K. " 'Linguistic Distance' as a Determinant of Bilateral Trade." *Southern Economic Journal* 72, no. 1 (2005): 1–15.

Isphording, Ingo E., and Sebastian Otten. "Linguistic Barriers in the Destination Language Acquisition of Immigrants." *Journal of Economic Behavior & Organization* 105 (2014): 30–50.

Iwanowsky, Mathias. "The Effects of Migration and Ethnicity on African Economic Development." Working Paper, 2018.

Jacks, David S., Christopher M. Meissner, and Dennis Novy. "Trade Costs, 1870–2000." *American Economic Review* 98, no. 2 (2008): 529–34.

Jain, Tarun. "Common Tongue: The Impact of Language on Educational Outcomes." *Journal of Economic History* 77, no. 2 (2017): 473–510.

Jia, Ruixue. "Weather Shocks, Sweet Potatoes and Peasant Revolts in Historical China." *Economic Journal* 124, no. 575 (2014): 92–118.

Kessinger, Tom G. "Regional Economy (1757–1857): North India." *Cambridge Economic History of India* 2 (1983): 242–70.

Kiszewski, Anthony, Andrew Mellinger, Andrew Spielman, Pia Malaney, Sonia Ehrlich Sachs, and Jeffrey Sachs. "A Global Index Representing the Stability of Malaria Transmission." *American Journal of Tropical Medicine and Hygiene* 70, no. 5 (2004): 486–98.

Krishnamurti, Bhadriraju. *The Dravidian Languages*. Cambridge: Cambridge University Press, 2003.

Kumar, Dharma. "Regional Economy (1757–1857): South India." *Cambridge Economic History of India* 2 (1983): 352–75.

Laitin, David, and Rajesh Ramachandran. "Language Policy and Human Development." *American Political Science Review* 110, no. 3 (2016): 457–80.

Lameli, Alfred, Volker Nitsch, Jens Südekum, and Nikolaus Wolf. "Same Same But Different: Dialects and Trade." *German Economic Review* 16, no. 3 (2015): 290–306.

Laval, Guillaume, Etienne Patin, and Valeria Rueda. "Achieving the American Dream: Cultural Distance, Cultural Diversity and Economic Performance." Oxford Economic and Social History Working Paper No. 140, Oxford, UK, 2016.

Markovits, Claude. *Merchants, Traders, Entrepreneurs: Indian Business in the Colonial Era*. London, UK: Springer, 2008.

Matsuura, Kenji, and Cort Willmott. "Terrestrial Air Temperature and Precipitation: 1900–2006 Gridded Monthly Time Series, Version 1.01." University of Delaware, Newark, Delaware, 2007.

McAlpin, Michelle. "Railroads, Prices, and Peasant Rationality: India 1860–1900." *Journal of Economic History* 34, no. 3 (1974): 662–84.

———. "Price Movements and Economic Activity (1860–1947)." *Cambridge Economic History of India* 2 (1983): 878–904.

Melitz, Jacques, and Farid Toubal. "Native Language, Spoken Language, Translation and Trade." *Journal of International Economics* 93, no. 2 (2014): 351–63.

Michalopoulos, Stelios. "The Origins of Ethnolinguistic Diversity." *American Economic Review* 102, no. 4 (2012): 1508–39.

Montaut, Annie. "Colonial Language Classification, Post-Colonial Language Movements and the Grassroot Multilingualism Ethos in India." *Mushirul Hasan & Asim Roy. Living Together Separately. Cultural India in History and Politics* (2005): 75–116.

Moxham, Roy. *The Great Hedge of India*. London, UK: Constable, 2001.

Nunn, Nathan, and Diego Puga. "Ruggedness: The Blessing of Bad Geography in Africa." *Review of Economics and Statistics* 94, no. 1 (2012): 20–36.

Nunn, Nathan, and Leonard Wantchekon. "The Slave Trade and the Origins of Mistrust in Africa." *American Economic Review* 101, no. 7 (2011): 3221–52.

O'Rourke, Kevin H., and Jeffrey G. Williamson. "When Did Globalisation Begin?" *European Review of Economic History* 6, no. 1 (2002): 23–50.

Özak, Ömer. "The Voyage of Homo-Economicus: Some Economic Measures of Distance." Working Paper, Department of Economics, Southern Methodist University, Dallas, TX, 2010.

———. "Distance to the Technological Frontier and Economic Development." *Journal of Economic Growth* 23, no. 2 (2018): 175–221.

Pandit, Prabodh Bechardas. *Language in a Plural Society*. New Delhi: Dev Raj Chanana Memorial Committee, 1977.

Pascali, Luigi. "The Wind of Change: Maritime Technology, Trade and Economic Development." *American Economic Review* 107, no. 9 (2017): 2821–54.

Pemberton, Trevor J., Michael DeGiorgio, and Noah A. Rosenberg. "Population Structure in a Comprehensive Genomic Data Set on Human Microsatellite Variation." *G3: Genes, Genomes, Genetics* 3, no. 5 (2013): 891–907.

Persaud, Alexander. "Escaping Local Risk by Entering Indentureship: Evidence from NineteenthCentury Indian Migration." *Journal of Economic History* 79, no. 2 (2019): 447–76.

Persson, Karl Gunnar. *Grain Markets in Europe, 1500–1900: Integration and Deregulation*. Vol. 7. Cambridge: Cambridge University Press, 1999.

Ramankutty, Navin, Jonathan A. Foley, John Norman, and Kevin McSweeney. "The Global Distribution of Cultivable Lands: Current Patterns and Sensitivity to Possible Climate Change." *Global Ecology and Biogeography* 11, no. 5 (2002): 377–92.

Rauch, James E., and Vitor Trindade. "Ethnic Chinese Networks in International Trade." *Review of Economics and Statistics* 84, no. 1 (2002): 116–30.

Renfrew, Colin. "The Origins of Indo-European languages." *Scientific American* 261, no. 4 (1989): 106–15.

Richards, John F. *The Mughal Empire*. Vol. 5. Cambridge: Cambridge University Press, 1995.

Roy, Tirthankar. *India in the World Economy: From Antiquity to the Present*. Cambridge: Cambridge University Press, 2012.

———. "Trading Firms in Colonial India." *Business History Review* 88, no. 1 (2014): 9–42.

Shastry, Gauri Kartini. "Human Capital Response to Globalization Education and Information Technology in India." *Journal of Human Resources* 47, no. 2 (2012): 287–330.

Shiue, Carol H., and Wolfgang Keller. "Markets in China and Europe on the Eve of the Industrial Revolution." *American Economic Review* 97, no. 4 (2007): 1189–216.

Spolaore, Enrico, and Romain Wacziarg. "The Diffusion of Development." *Quarterly Journal of Economics* 124, no. 2 (2009): 469–529.

——— . "Fertility and Modernity." UCLA CCPR Population Working Papers No. PWP-CCPR-2016016, Los Angeles, CA, 2016.

———. "Ancestry and Development: New Evidence." *Journal of Applied Econometrics* 33, no. 5 (2018): 748–62.

Studer, Roman. "India and the Great Divergence: Assessing the Efficiency of Grain Markets in Eighteenth-and Nineteenth-Century India." *Journal of Economic History* 68, no. 2 (2008): 393–437.

Waldinger, Maria. "The Economic Effects of Long-Term Climate Change: Evidence from the Little Ice Age." Working Paper, London School of Economics, London, UK, 2014.

Weir, Bruce S., and C. Clark Cockerham. "Estimating F-Statistics for the Analysis of Population Structure." *Evolution* 38, no. 6 (1984): 1358–70.

Wichmann, Søren, Eric W. Holman, and Cecil H. Brown (eds.). "The ASJP Database (Version 17)," https://doi.org/10.5281/zenodo.3835942, 2016.